AGU PUBLICATIONS

JGR

Journal of Geophysical Research: Solid Earth

RESEARCH ARTICLE

10.1002/2014JB011777

Key Points:
- We apply statistical tests to synthetic clustered earthquake catalogs
- Success of a test depends on the test, type of clustering, and quantity of data
- The fraction of triggered global $M = 7$ events is below 20% for aftershocks

Correspondence to:
E. G. Daub,
egdaub@memphis.edu

Citation:
Daub, E. G., D. T. Trugman, and P. A. Johnson (2015), Statistical tests on clustered global earthquake synthetic data sets, *J. Geophys. Res. Solid Earth*, *120*, 5693–5716, doi:10.1002/2014JB011777.

Received 14 NOV 2014
Accepted 23 JUL 2015
Accepted article online 31 JUL 2015
Published online 22 AUG 2015

Statistical tests on clustered global earthquake synthetic data sets

Eric G. Daub[1], Daniel T. Trugman[2,3], and Paul A. Johnson[2]

[1]Center for Earthquake Research and Information, University of Memphis, Memphis, Tennessee, USA, [2]Geophysics Group, Los Alamos National Laboratory, Los Alamos, New Mexico, USA, [3]Scripps Institution of Oceanography, University of California, San Diego, La Jolla, California, USA

Abstract We study the ability of statistical tests to identify nonrandom features of earthquake catalogs, with a focus on the global earthquake record since 1900. We construct four types of synthetic data sets containing varying strengths of clustering, with each data set containing on average 10,000 events over 100 years with magnitudes above $M = 6$. We apply a suite of statistical tests to each synthetic realization in order to evaluate the ability of each test to identify the sequences of events as nonrandom. Our results show that detection ability is dependent on the quantity of data, the nature of the type of clustering, and the specific signal used in the statistical test. Data sets that exhibit a stronger variation in the seismicity rate are generally easier to identify as nonrandom for a given background rate. We also show that we can address this problem in a Bayesian framework, with the clustered data sets as prior distributions. Using this new Bayesian approach, we can place quantitative bounds on the range of possible clustering strengths that are consistent with the global earthquake data. At $M = 7$, we can estimate 99th percentile confidence bounds on the number of triggered events, with an upper bound of 20% of the catalog for global aftershock sequences, with a stronger upper bound on the fraction of triggered events of 10% for long-term event clusters. At $M = 8$, the bounds are less strict due to the reduced number of events. However, our analysis shows that other types of clustering could be present in the data that we are unable to detect. Our results aid in the interpretation of the results of statistical tests on earthquake catalogs, both worldwide and regionally.

1. Introduction

The occurrence of a number of earthquakes above magnitude $M = 8$ in the past decade has led to speculation that large earthquakes cluster in time. The recent occurrence of earthquakes includes three of the six largest earthquakes on record, including the 2004 $M_W = 9.2$ Sumatra earthquake, the 2010 $M_W = 8.8$ Maule, Chile, earthquake, and the 2011 $M_W = 9.1$ Tohoku earthquake, and a number of additional events above $M = 8$, as can be seen from the earthquake record shown in Figure 1a. These large events can have an outsized effect on seismic hazard through their large release of stored elastic energy, causing destructive strong ground motions and tsunamis. If large events do cluster in time, this could change the way that seismic hazard is estimated worldwide.

To evaluate this hypothesis, a number of studies have compared the global earthquake record since 1900 to a process that is random in time [*Bufe and Perkins*, 2005; *Michael*, 2011; *Shearer and Stark*, 2012; *Daub et al.*, 2012; *Parsons and Geist*, 2012; *Ben-Naim et al.*, 2013]. A process that is random in time is also known as a time-homogeneous Poisson process, and such a process assumes that the event occurrence times are uncorrelated. The majority of these studies tends to show that earthquake occurrence worldwide since 1900 shows no deviation from a process that is random in time, other than localized aftershock sequences. This is illustrated in Figure 1b for the *Ben-Naim et al.* [2013] study, showing the likelihood that the catalog is random (through a p value, calculated using Monte Carlo simulation) for several magnitude thresholds with and without aftershock removal. While some of the statistical tests applied to the catalog appear to show deviations from random event occurrence at minimum magnitude levels $M = 8.4–8.6$ [*Bufe and Perkins*, 2005; *Ben-Naim et al.*, 2013], such as in Figure 1 at $M \geq 8.4–8.5$, these deviations are not strong enough to conclude that the earthquake record is nonrandom. This is because tests using magnitude thresholds selected a posteriori underestimate p values, as shown by *Shearer and Stark* [2012]. The global catalog over this time period is regarded as complete only for $M \geq 7$; thus, the catalogs in these studies contain a relatively small number of events, particularly at high-magnitude levels.

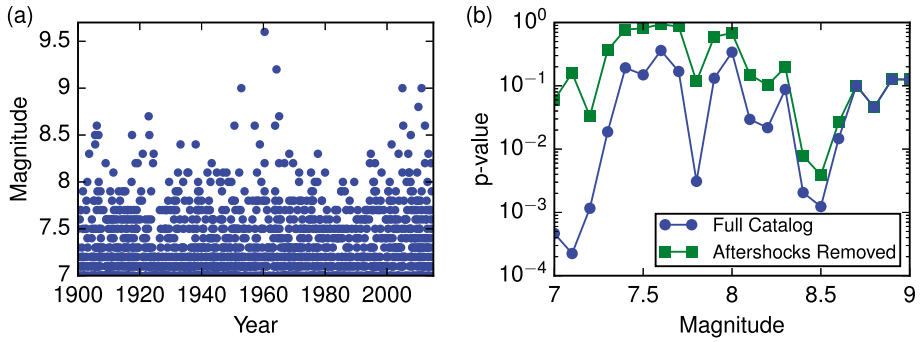

Figure 1. (a) PAGER Global earthquake catalog from 1900 to 2014 [*Allen et al.*, 2009], supplemented by the United States Geological Survey preliminary determination of epicenters catalog through the end of 2014. Occurrence of several earthquakes with $M \geq 8.5$ from 1950–1964 and 2004–2012 led to speculation that the largest earthquakes cluster in time. (b) Statistical analysis using the variance of the recurrence time [*Ben-Naim et al.*, 2013], showing calculated p values as a function of threshold magnitude, both with and without removal of aftershocks. The p value is determined by Monte Carlo simulation by calculating the fraction of random catalogs that exhibit a normalized variance that exceeds the value calculated for the PAGER catalog. The results show that, in general, the earthquake record does not deviate from a process that is random in time. For magnitude levels $M \geq 8.4$–8.5, the p values appear to be low, though they are not low enough to conclude that the earthquake record is nonrandom, as the reported p values are underestimates due to selection of the magnitude levels post hoc [*Shearer and Stark*, 2012].

More recent seismic catalogs have been used to examine the ability of large earthquakes to trigger earthquakes above $M = 5$. One study showed that the 2012 $M_w = 8.6$ Indian Ocean event triggered aftershocks above $M = 5$ worldwide [*Pollitz et al.*, 2012] followed by a quiescent period [*Pollitz et al.*, 2014]. A more comprehensive study looking at many events above $M = 7$ showed that triggering of this nature may not be common [*Parsons and Velasco*, 2011], while another more recent study concluded that there is no evidence for elevated seismicity rates from $M = 5.2$ to 5.6 over the recent years [*Parsons and Geist*, 2014]. Further, the *Parsons and Velasco* [2011] study showed that instantaneous triggering (i.e., triggering coincident with surface wave arrivals) of larger events above $M = 5$ is not observed as frequently as observed triggering rates for smaller $M < 5$ earthquakes would predict [*Velasco et al.*, 2008]. The observational evidence thus suggests that small events are routinely triggered [*Hill et al.*, 1993; *Gomberg et al.*, 2004; *Freed*, 2005], and there is some evidence that large earthquakes can have global effects on moderate-sized events. However, the likelihood that large events trigger other moderately large events on a global scale remains unclear.

In this study, we address the ability of statistical tests to identify catalogs as nonrandom, with a particular focus on the global data set since 1900. We produce a series of simulated data sets that are clustered by construction and systematically vary the strength of the clustering to assess how well different statistical tests can identify these data sets as nonrandom. While this question has been addressed previously in a study by *Dimer de Oliveira* [2012], here we perform a more systematic study of various data set types with clustering of different strengths. In particular, our study aims to bound the range of clustering strengths that are most likely to be consistent with the global earthquake catalog. Through these synthetic tests, we can assess the results of the numerous studies performed on the global earthquake record and interpret the implications for earthquake interaction and earthquake hazard worldwide.

2. Synthetic Data Sets

We generate four types of synthetic data sets that are nonrandom by construction for analysis using various statistical tests. The synthetic data sets are each designed to be similar to the global earthquake record since 1900, with a few simplifications. The simulations are all 100 years in length and contain an average of 10,000 events above a minimum magnitude threshold of $M = 6$, illustrated in Figure 2. Event magnitudes are drawn from a Gutenberg-Richter distribution [*Gutenberg and Richter*, 1954], with a cumulative distribution function $\mathrm{CDF}(M) \propto 10^{-bM}$ with $b = 1$, a minimum magnitude of $M_{\mathrm{min}} = 6$, and a maximum magnitude of $M_{\mathrm{max}} = 9.5$, with a few additional simulations varying these parameters described below. The magnitude-frequency statistics of this distribution are shown in Figure 3. This distribution shows that the data sets contain ~10,000 events at $M \geq 6$, ~1000 events with $M \geq 7$, and ~100 events with $M \geq 8$. All of these numbers differ slightly from the observed values in the earthquake record since 1900—the Prompt Assessment of Global Earthquakes

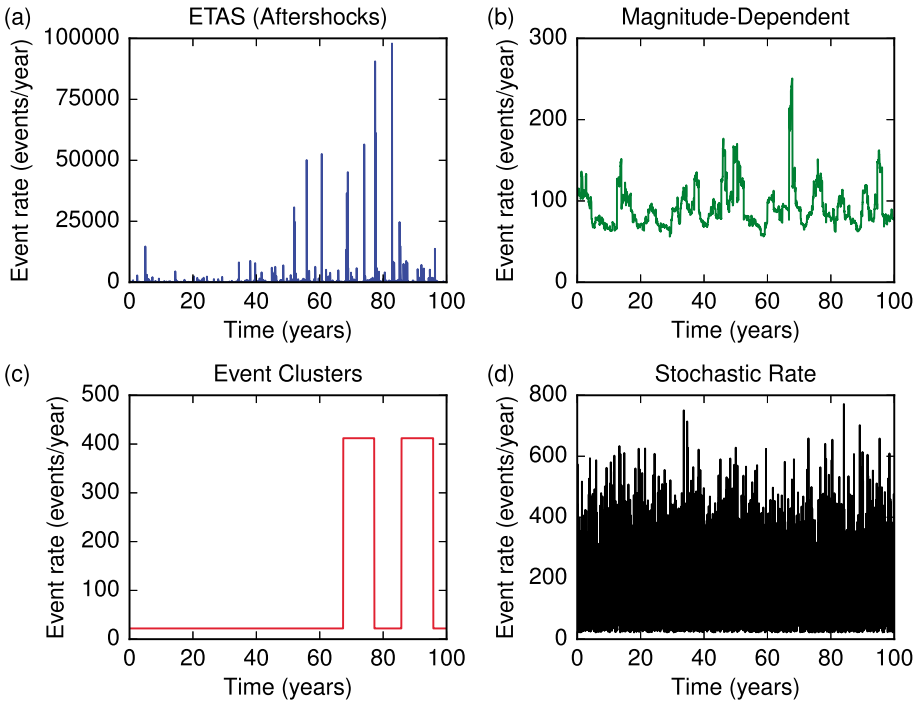

Figure 2. Examples of synthetic clustered data sets. All plots show earthquake event rate as a function of time for a 100 year sequence containing approximately 10,000 events above $M = 6$. The plots shown here illustrate the most strongly clustered example of each simulation type. (a) ETAS model, where additional events are added to the synthetic data set following empirical aftershock rules. The seismicity rate is sharply peaked following large events, decaying rapidly in time. (b) Magnitude-dependent rate, where the seismicity rate changes based on the magnitude of the previous 200 events. (c) Event clusters, where two 10 year clusters of events are placed randomly in time. (d) Stochastic rate simulations, where the rate varies stochastically in time from event to event, with the variations following a one-sided normal distribution.

for Response (PAGER) catalog [*Allen et al.*, 2009] supplemented with the United States Geological Survey preliminary determination of epicenters (PDEs) catalog contains ~1800 events above $M = 7$ over 115 years.

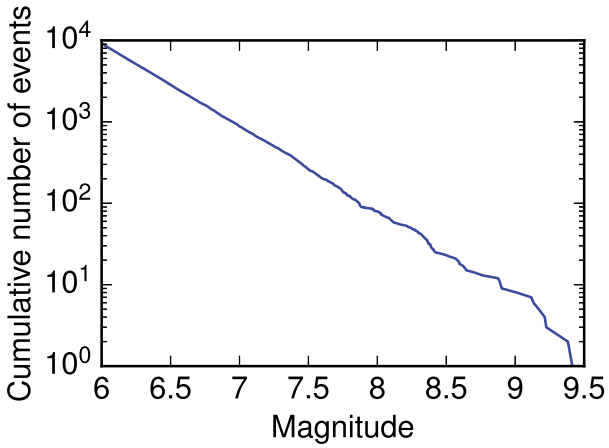

Figure 3. Magnitude-frequency distribution used in the synthetic data sets. Events follow a Gutenberg-Richter distribution with $b = 1$ and a minimum magnitude of $M_{min} = 6$. While this differs from the empirical value of $b = 1.26$ observed for the PAGER/PDE catalog, using $b = 1$ makes determining the number of events at a given magnitude level more straightforward. Alternative magnitude-frequency distributions with $b = 0.8$, $b = 1.2$, and $M_{min} = 5$ are also considered for the ETAS simulations.

The combined PAGER and PDE catalog is estimated to be complete for $M \geq 7$ using both the maximum curvature and b value stability methods [*Woessner and Wiemer*, 2005]. With this completeness magnitude, $b = 1.26$ in the observed magnitude-frequency distribution using the maximum likelihood method [*Aki*, 1965]. However, an analysis of the more recent International Seismological Centre Global Earthquake Model catalog [*Storchak et al.*, 2013] shows that for events in that catalog since 1917, $b \approx 1$ [*Michael*, 2014]. For the PAGER global catalog, the number of independent events above $M = 7$ varies from ~500 to 1500, depending on the criteria used in aftershock identification [*Michael*, 2011; *Shearer and Stark*, 2012; *Daub et al.*, 2012]. Because a different choice of declustering algorithm can always alter the number of events in the catalog, for simplicity we use $b = 1$ with a

rate of 100 events/year to produce a round number of events at various magnitude levels for the majority of our simulations. To evaluate if our results are sensitive to the details of the magnitude distribution, we also perform additional simulations with different b values of 0.8 and 1.2 and a set of simulations with a minimum magnitude of $M_{\text{min}} = 5$.

For each type of data set, we create 20 different versions, varying the clustering strength. The background rate ranges from 98 events/year in the least clustered version to 22 events/year in the most clustered version. According to each prescription below, we either explicitly add additional events according to certain rules (the case for the Epidemic Time Aftershock Sequence (ETAS) simulations) or vary the seismicity rate from event to event using certain rules (the case for the other three synthetic data sets). In each case, parameters are selected such that our simulations produce on average 100 events/year over a fixed 100 year duration. We produce 10,000 realizations for each of the 20 different clustering strengths of the four data set types. Thus, our results here are based on analyzing a total of 800,000 simulations (four clustering types × 20 clustering strengths × 10,000 realizations).

2.1. ETAS Simulations

Epidemic Time Aftershock Sequence (ETAS) models have been developed over many years to represent empirically observed features of aftershock sequences [*Ogata*, 1998]. Because of the widespread use of these types of simulations to represent clustering of seismicity, we generate a series of data sets containing "global aftershock sequences" to represent one potential type of clustering in our analysis. In the ETAS models, events are added as aftershocks of background events according to two empirical rules. First, an aftershock productivity law determines the number of aftershocks produced by a main shock of magnitude M [*Felzer et al.*, 2004; *Helmstetter et al.*, 2005]:

$$N_{\text{AS}} = C' 10^{\alpha(M - M_{\text{min}})}. \tag{1}$$

The constant α determines the relative number of aftershocks an earthquake of a given magnitude produces, and C', known as the aftershock productivity, determines the overall number. Observations show that $\alpha \approx 1$, and C' is such that a main shock typically produces a maximum aftershock magnitude one magnitude unit less than the main shock (Båth's law) [*Helmstetter and Sornette*, 2003]. Here we use $\alpha = 1$ but vary the aftershock productivity to produce simulations with different levels of clustering.

The second empirical rule for the ETAS models is the Omori decay of aftershock rate with time following the main shock [*Omori*, 1895; *Utsu et al.*, 1995]:

$$R(t) = \frac{A}{(c + t)^{\beta}}. \tag{2}$$

In the Omori law, β describes the time decay of aftershock activity, and c is a constant to make the rate finite at $t = 0$. We use $\beta = 1.07$ and $c = 3 \times 10^{-4}$ years = 0.11 days, which are similar to ETAS parameters for Japan [*Guo and Ogata*, 1997]. The constant A is chosen such that equation (2) integrates to N_{AS}:

$$N_{\text{AS}} = \int_0^{t_{\text{max}}} R(t) \mathrm{d}t. \tag{3}$$

We cut off the aftershock decay after $t_{\text{max}} = 100$ years since that would exceed the length of our simulation. This truncation of the aftershock sequence is strictly necessary only if $\beta \leq 1$.

Aftershocks in the ETAS model also produce their own aftershocks, so equations (1) and (2) are applied recursively to all aftershocks until no further aftershocks are produced. Note that aftershocks follow the same magnitude-frequency statistics as the main shocks, which means that most aftershocks are small events. However, occasionally, an event triggers an aftershock whose magnitude exceeds the parent event. In this case, the initial event is considered to be a foreshock. A consequence of this is that not all sets of ETAS parameters produce aftershock sequences that terminate [*Helmstetter and Sornette*, 2002], and so care must be taken in selecting parameters.

Because the ETAS models are event based, we set the background rate to the desired level and then choose C' such that the synthetic data sets contain a total of 10,000 events in 100 years on average. Specific values of the rate and the aftershock productivity C' are shown in Table 1. Because aftershock sequences can extend

Table 1. Parameter Values for the Synthetic Data Sets Used in the Study[a]

Branching Ratio	Background Rate (Events/Year)	ETAS C'	Magnitude γ (Events/Year)	Clusters λ_{clust} (Events/Year)	Stochastic σ (Events/Year)
0.02	98	0.0027	0.001	10	3
0.06	94	0.0076	0.004	30	8
0.1	90	0.0125	0.0075	50	14
0.14	86	0.0175	0.01	70	19
0.18	82	0.0225	0.0135	90	25
0.22	78	0.0275	0.0165	110	31
0.26	74	0.0325	0.02	130	38
0.3	70	0.037	0.023	150	45
0.34	66	0.0425	0.026	170	53
0.38	62	0.047	0.03	190	60
0.42	58	0.0525	0.0325	210	68
0.46	54	0.057	0.0365	230	77
0.5	50	0.0623	0.04	250	87
0.54	46	0.06725	0.0435	270	97
0.58	42	0.072	0.047	290	108
0.62	38	0.07725	0.0515	310	120
0.66	34	0.0825	0.0545	330	135
0.7	30	0.08725	0.0585	350	150
0.74	26	0.0925	0.062	370	165
0.78	22	0.0975	0.066	390	185

[a]Details of all models are described in the main text.

over long periods of time for the highest values of aftershock productivity, we ensure that the synthetic data sets are uniform in time by generating 200 years of background events and only selecting the final 100 years of the event sequence for analysis. This ensures that we are not "missing" events whose main shock may have occurred prior to the start of the simulation.

ETAS models can also include spatial kernels to simulate clustering of seismicity in space [*Felzer and Brodsky*, 2006]. Because most statistical tests applied to the global catalog neglect spatial information, we do not include this effect in our simulations. However, spatial information is implicitly included in statistical tests on the global catalog through removal of aftershocks, so a more realistic way to treat the ETAS data for our purposes might be to perform a spatial ETAS simulation and then remove aftershocks prior to testing. Due to the large number of simulations considered here and the introduction of additional parameters for both generating and removing aftershocks, we neglect this aspect in our study. Additionally, we note that our simulations approximate this spatial effect, due to the fact that the Omori decay of the aftershock rate with time is assumed to be independent of spatial location. Thus, an ETAS simulation that uses a reduced aftershock productivity can be thought of as a proxy for a spatial simulation where events in the traditional aftershock zone are removed. This leaves a reduced number of global "aftershocks" in the simulated data set that follow the same time decay of seismicity rate as the traditional aftershocks.

A sample ETAS simulation for the strongest level of clustering is shown in Figure 2a, illustrating the seismicity rate as a function of time. Because large earthquakes have a pronounced effect on the rate (equations (1)–(2)), a localized spike in the earthquake rate occurs after each large event. The rate quickly decays with time after each event, though extended aftershock sequences can occur that keep the rate elevated above background for longer periods of time. Due to the sharp peak in the rate following large events, the ETAS models exhibit a higher variability in rate when compared to the other simulated data sets in this study.

2.2. Magnitude-Dependent Simulations

The second type of data set incorporates magnitude-dependent clustering, designed to be somewhat similar to the ETAS models yet different in its time dependence. It is based on the idea that earthquake triggering is related to the strain amplitude of seismic waves, combined with a model where any earthquake can be a potential precursor to future seismicity with a rate contribution dependent on magnitude [*Rhoades and Evison*, 2004]. In particular, a study by *van der Elst and Brodsky* [2010] showed that the seismicity rate increase could be quantitatively tied to wave strain amplitude using a statistical method applied to a large earthquake catalog. We construct a synthetic data set based on this concept, where the seismicity rate at a given time is the sum of the background rate, plus a variable contribution that depends on the weighted magnitude of the previous 200 events. The seismicity rate λ_i for the time period following event i is given by

$$\lambda_i = \lambda_0 + \gamma \left[\sum_{j=i-199}^{i} 10^{\alpha(M_j - M_{\min})} - 200 \right]. \quad (4)$$

The background seismicity rate is λ_0, $\alpha = 1$ determines the relative contribution of earthquakes of different magnitude in the same manner as the ETAS models, and γ determines the overall rate contribution from the triggering effect. Seismic wave amplitudes scale exponentially with magnitude as $\sim 10^M$ [*Lay and Wallace*, 1995]; thus, we apply a weighting factor that depends exponentially on magnitude to determine their strain amplitude contribution. As with the ETAS models, spatial information is neglected. Finally, subtracting the factor of 200 ensures that the rate increase is measured relative to a baseline level where all events have magnitude $M = 6$. Following a large event, the seismicity rate exhibits an approximate step increase and remains elevated for a longer period of time when compared to the ETAS simulations (Figures 2a and 2b). The additive combination of past events on top of the background seismicity is the same as in the model of *Rhoades and Evison* [2004], though with a cutoff after a limited number of events.

The duration over which far-field dynamic triggering occurs is not well quantified by observations, as event statistics are usually accumulated over longer time periods to establish a change in seismicity rate [*Freed*, 2005]. Some studies suggest that near-field aftershocks may also be triggered by dynamic stresses [*Gomberg et al.*, 2003], suggesting an Omori time dependence for the triggering effect. Because the ETAS simulations already consider the case of clustering following an Omori decay of seismicity rate with time, we use the magnitude-dependent simulations to consider an alternative form of a seismicity rate increase that is not as pronounced in time. *van der Elst and Brodsky* [2010] show that a step rate change following the triggering event predicts similar triggering statistics to an Omori decay, and thus, we use equation (4) to generate this rate pattern in this simulated data set. The convolution over the previous 200 magnitudes is aimed at making this step change last from 1 to 2 years, depending on the size of the rate increase. The choice of 200 events (rather than a specific period of time) is also made to ensure that events with $M \geq 8$ influence the seismicity rate over multiple $M \geq 8$ recurrence times (since 1 out of every 100 events will have a magnitude in that range).

For the magnitude-dependent simulations, we generate 15,200 events with magnitudes following the Gutenberg-Richter distribution and then perform the sum over event magnitudes to determine the event-by-event seismicity rate (equation (4)). We then generate recurrence times in an event-by-event fashion by drawing from an exponential distribution (the expected waiting time distribution for a Poisson process) given λ_i. Finally, we sum the recurrence times to get the occurrence times and remove the first 200 events to ensure uniformity in time. We then select 100 years of events for analysis and discard events that occur after 100 years. We produce 5000 additional events beyond the typical 10,000 to ensure that our simulations are never shorter than 100 years in the event of a large number of short interevent times. The coefficient γ is determined for each of the 20 synthetic data sets in order to produce simulations that contain on average 10,000 events in 100 years. Parameter values are shown in Table 1.

An example magnitude-dependent synthetic data set is shown in Figure 2b for the highest clustering strength. When compared to the ETAS model, the spikes in the seismicity rate for the magnitude-dependent simulations are much reduced in amplitude and are extended in time. This is because the rate increases are much less localized in time when compared to an Omori decay—when a large event occurs, the seismicity rate is elevated but fairly constant over the ensuing time period.

The relative lack of changes in the seismicity rate for the magnitude-dependent simulations illustrates an important issue in earthquake statistics. Because the time period over which the effects of an earthquake are

evident is reasonably long, the true background seismicity rate (22 events/year) is rarely observed in Figure 2b. There is always some earthquake that is influencing the seismicity rate, causing the nominal background rate (i.e., the rate inferred by an observer through examination) to not reflect the true background rate. While this problem exists in all earthquake records (observed and simulated), its effect is more pronounced here because the true background rate is never observed. This type of clustering is therefore more difficult to detect, despite the true background rate being identical to the other data sets, because the tests cannot distinguish the nominal and true background rates.

2.3. Event Clusters

We generate a third type of nonrandom data set by inserting two randomly placed, nonoverlapping 10 year clusters with an elevated seismicity rate. In these simulations, the seismicity rate is elevated from λ_0 to a level $\lambda_0 + \lambda_{\text{clust}}$ for a specified number of events N_{clust} following two random events, chosen such that the two clusters are nonoverlapping and that both clusters fit within the total length of the simulation. The number of events with an elevated rate N_{clust} increases as the clustering strength increases. The number of events N_{clust} and the rate change λ_{clust} are chosen so that the duration of each cluster is on average 10 years, and the data set contains on average 10,000 events over its 100 year duration. Given a desired background rate λ_0, one can determine N_{clust} and λ_{clust} by

$$N_{\text{clust}} = \frac{10,000 - \lambda_0 \times 100\text{years}}{2} + \lambda_0 \times 10\text{years} \tag{5}$$

$$\lambda_{\text{clust}} = \frac{10,000 - \lambda_0 \times 100\text{years}}{2 \times 10\text{years}}. \tag{6}$$

The particular values of λ_{clust} given our values of λ_0 can be found in Table 1. As with the magnitude-dependent rate simulations, the simulations are created in practice by drawing 15,000 recurrence times from an appropriately scaled exponential distribution given the seismicity rate for a particular event and then restricting our analysis to 100 years. As with the magnitude-dependent data sets, we produce far more than 10,000 events to be certain that all of our simulations are 100 years in duration. Magnitudes are assigned independent of the seismicity rate, following the same distribution as the other simulations.

This model is inspired by the "clusters" of magnitude 8.4 and larger events observed in the years 1950–1962 and 2004–2012 [*Bufe and Perkins*, 2005; *Ben-Naim et al.*, 2013]. While there is no clear evidence that seismicity at lower levels was higher during these time periods, we use this aspect of the global earthquake catalog as a guide in creating our synthetic data sets to better understand how strong such clusters would need to be in order to be detectable by statistical tests. We also note that the study by *Dimer de Oliveira* [2012] used temporal clusters of larger ($M \geq 8$) events in his analysis of simulated clustered data, giving us a point of comparison between our study and other work on this topic. An example of the seismicity rate as a function of time for an event clusters simulation is shown in Figure 2c.

2.4. Stochastic Rate Simulations

The final type of data set that we consider is one in which the rate varies stochastically in time. As with the magnitude-dependent and event clusters simulations, the rate λ_i changes from event to event. For the stochastic rate data sets, the event rate following event i is $\lambda_i = \lambda_0 + \lambda_{\sigma,i}$, where $\lambda_{\sigma,i}$ is drawn from a one-sided normal distribution. The distribution peaks at zero (i.e., the equivalent two-sided normal distribution has zero mean) and has a standard deviation σ. To set the level of clustering, we vary the value of σ. As with the other simulations, σ is chosen given λ_0 so that the 100 year portion of the synthetic data set that we use in the analysis contains on average 10,000 events. The specific values for each background rate are presented in Table 1.

This set of simulations was selected to provide a scenario where the rate does not vary in a predictable fashion with time, and consequently the nonrandom character of the simulation is very difficult to detect, as we will see through our analysis. An example of a Stochastic Rate simulation with the strongest level of clustering is shown in Figure 2d.

3. Quantifying Clustering Strength

To calibrate the parameter values used in generating our synthetic data and to quantify the clustering in the resulting event sequences, we use two different measures of the clustering strength. The first is the branching ratio, defined as the fraction of events that are in excess of the background seismicity level [*Sornette and*

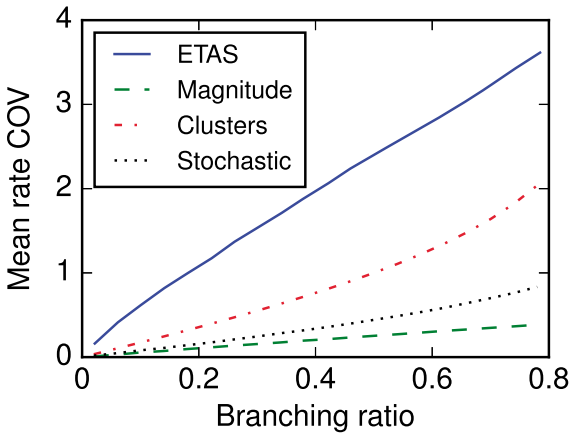

Figure 4. Mean coefficient of variation for the event rate over all synthetic data sets as a function of branching ratio (or fraction of the events that are not background events) for the four synthetic data sets considered in this study. In each case, there is an approximately linear relationship between branching ratio and mean rate COV, and this illustrates that we have a quantitative measure of the clustering strength for each synthetic data set. Some clustering types exhibit larger rate variations for the same branching ratio when compared to others, and our results show that data sets with stronger rate variations are generally easier to detect as nonrandom.

Sornette, 1999]. This quantity was initially developed for ETAS models, as it represents the fraction of events that are aftershocks, but it applies generally to all of the models considered here through our knowledge of the background seismicity rate. For each synthetic data set, we generate 20 different random ensembles with branching ratios varying from 0.02 to 0.78 with a spacing of 0.04. This means that the background rate ranges from 98 events/year to 22 events/year, with a difference of 4 events/year between each successive clustering strength. The parameter controlling clustering strength for each simulation type is then changed by trial and error so that the simulations show an overall event rate of 100 events/year. Parameter values for the different models are provided in Table 1.

While the branching ratio is only dependent on the background event rate, the actual details of how the event rate varies with time change across the different synthetic data sets. Thus, we would like to have an independent measure based on fluctuations in the rate that quantifies the strength of clustering. For each type of synthetic data set, we calculate the coefficient of variation (COV, the standard deviation normalized by the mean) of the rate, assuming the rate is piecewise constant in time. Then we find the mean over all 10,000 realizations (only 1000 realizations are used in the rate calculations for the ETAS simulations, due to a greater computational cost associated with determining the rate in the ETAS model). The mean rate COV is shown as a function of branching ratio for all four types of synthetic data sets in Figure 4. For each type, the mean rate COV is linearly dependent on the branching ratio, with a different constant of proportionality for each type. Because the details of the time variation in the rate change for each type of clustering, we do not expect the mean rate COV to be the same for each type of synthetic data. In particular, we find that the ETAS models have the most strongly varying rate, while the magnitude-dependent rate fluctuates the least in time.

In the remainder of this study, we use the branching ratio as our primary indicator of clustering strength due to its simple origin, ease of calculation, and uniformity across the different types of synthetic data sets. However, our results show that mean rate COV is a better indicator of how easily the nonrandom character of a particular data set can be detected, suggesting that rate variations give a better characterization of the nonrandom behavior for a particular type of synthetic data.

4. Statistical Tests

We select eight statistical tests used in the literature to test the synthetic data sets for deviations from random event occurrence. These tests can be divided into two classes: parameter-free tests and parameter-based tests. We choose four versions of each and apply each of them to all 800,000 realizations of our synthetic data sets. The parameter-free tests include a test based on the variance of the recurrence times [*Ben-Naim et al.*, 2013], a Kolmogorov-Smirnov (KS) test that compares recurrence times to an exponential distribution [*Michael*, 2011], a KS test that compares event occurrence times to a uniform distribution [*Shearer and Stark*, 2012], and an autocorrelation test [*Michael*, 2011; *Parsons and Geist*, 2012]. We also choose four tests that require choosing parameter values, including a multinomial chi-square test [*Gardner and Knopoff*, 1974; *Shearer and Stark*, 2012], a Poisson dispersion test [*Shearer and Stark*, 2012; *Daub et al.*, 2012], an alternative chi-square test [*Brown and Zhao*, 2002; *Luen and Stark*, 2012], and a test that looks for a seismicity rate increase following large events [*Michael*, 2011]. Due to the large number of simulated data sets, we do not vary the parameter values.

Table 2. *p* Values for the Statistical Tests Used in This Study Applied to the PAGER/PDE Catalog Through the End of 2014[a]

Catalog	*N*	Var	KSE	KSU	AC	MC	PD	BZ	Big Event
$M = 7$	1814	6.0×10^{-4}	1.4×10^{-4}	1.6×10^{-5}	0.95	0.0060	3.0×10^{-4}	0.0019	0.0076
$M = 7.5$	462	0.14	0.053	0.11	0.41	0.14	0.17	0.088	0.24
$M = 8$	87	0.34	0.71	0.15	0.29	0.60	0.59	0.51	0.29
$M = 7$ (AS)	1369	0.062	0.40	1.1×10^{-7}	0.58	0.24	0.010	0.024	0.87
$M = 7.5$ (AS)	385	0.81	0.69	0.18	0.21	0.66	0.69	0.60	0.64
$M = 8$ (AS)	77	0.68	0.81	0.31	0.24	0.22	0.91	0.78	0.24

[a]The tests include variance (Var), KS exponential (KSE), KS uniform (KSU), autocorrelation (AC), multinomial chi-square (MC), Poisson dispersion (PD), Brown and Zhao chi-square (BZ), and big event triggering, and the details on each test are provided in the main text. The tests are applied to the catalog both with and without removal of aftershocks ((AS) denotes removal of aftershocks, using the method described in *Daub et al.* [2012]) at minimum magnitude levels of $M = 7$, 7.5, and 8.

Most of these tests have been applied to the global earthquake data set, as well as other types of earthquake catalogs. We summarize the results of application of these tests to the PAGER/PDE data set through the end of 2014 in Table 2. The tests are applied at minimum magnitudes of 7, 7.5, and 8 both with and without removal of aftershocks using the procedure described in *Daub et al.* [2012]. In general, the catalog does not show a significant deviation from random event occurrence once aftershocks are removed, other than an apparent long-term variation in the seismicity rate at $M = 7$ that has been attributed to differences in event magnitude estimation over time [*Daub et al.*, 2012]. The *p* values here are in some cases different from values reported in the literature. This is due principally to differences in the declustering methods and to a lesser extent to differences in the Monte Carlo methods used to estimate the *p* values—in particular, for computational reasons we condition on the observed rate in this study rather than the observed number of events. When these differences are accounted for, our *p* values are in agreement with previously reported values.

For all of the results that follow, we will use the criteria that a deviation at the 1% level constitutes a nonrandom result, meaning that the data set in question has a value of the test statistic that is larger than 99% of random realizations. This is smaller than the frequently used value of 5% often used in hypothesis testing, as well as the 2.3% value associated with a test statistic that is more than two standard deviations from the mean for a normal distribution, another common cutoff. While this lower cutoff is chosen for conservative purposes, we have performed our analysis using different cutoff values, both higher and lower. We find the relative frequency of detection between different tests, clustering strengths, and magnitude levels to be independent of the choice of cutoff value. The particular value of the detection power for a specific test, clustering level, and magnitude cutoff will change if a different significance value is selected, but the relative values and overall trends remain unchanged.

4.1. Variance Test

The variance test compares the normalized variance of the recurrence times to the normalized variance expected if the event times are uncorrelated [*Ben-Naim et al.*, 2013]. Given *N* events, we compute a sequence of $N - 1$ recurrence (or interevent) times t_r and compute the normalized variance *V*

$$V = \frac{\langle t_r^2 \rangle - \langle t_r \rangle^2}{\langle t_r \rangle^2} \tag{7}$$

where $\langle \cdot \rangle$ indicates the average. The normalized variance is expected to be nearly unity for a random sequence of events, though for data sets with small numbers of events, the distribution peaks at a smaller value [*Ben-Naim et al.*, 2013]. A data set is flagged as nonrandom if the normalized variance is significantly larger than its expected value (i.e., the variance test is one sided), which is determined by Monte Carlo simulation using 10,000 random realizations. The random realizations have a background rate of 100 events/year and last for 100 years, and the magnitude distribution is drawn from the same distribution as the simulated data sets. Note that due to the high computational cost of simulating all possible numbers of events, here we condition on the background rate and duration rather than on the number of events as was done by *Ben-Naim et al.* [2013]. The variance test compares the entire distribution to the expected distribution for random event occurrence (i.e., all recurrence times are considered in calculating *V*), weighing the long recurrence times more heavily. If a data set contains an excess of long recurrence times, the test flags the sequence as nonrandom.

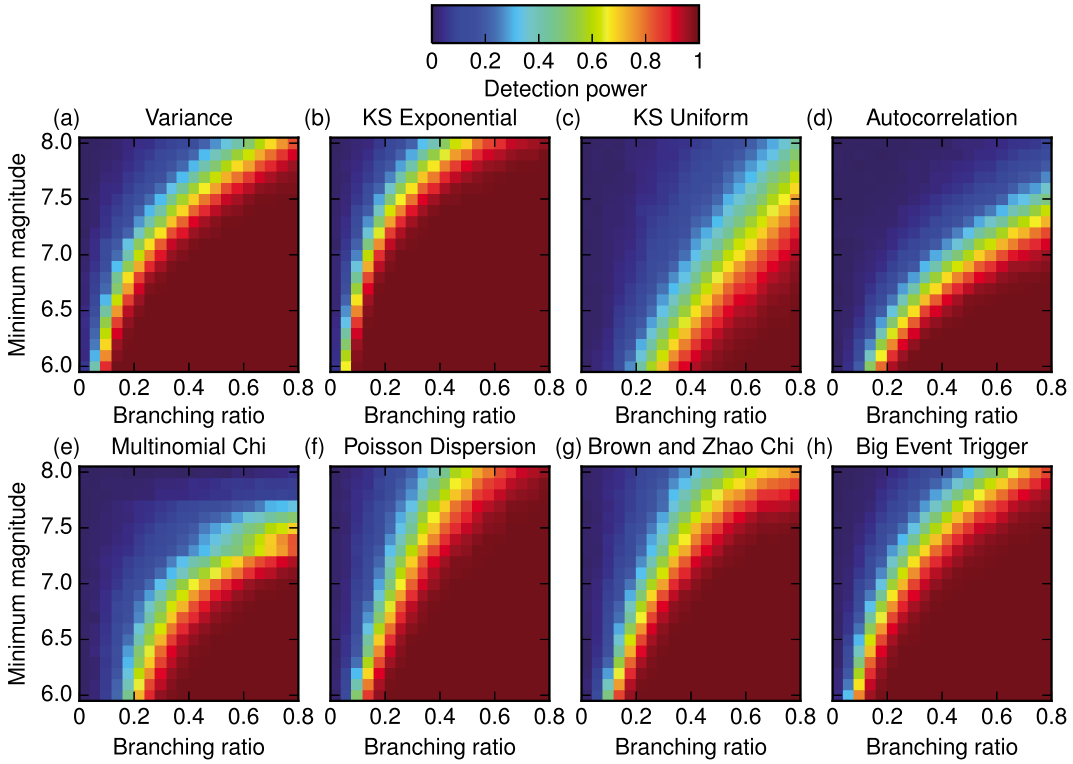

Figure 5. Detection power at the 1% level as a function of branching ratio and minimum magnitude for statistical tests applied to the ETAS models. Vertical axis indicates magnitude level, horizontal axis indicates branching ratio, and color scale indicates detection power or the probability that the test identifies the event sequence as nonrandom. Results are shown for a set of eight statistical tests: (a) variance test, (b) KS exponential test, (c) KS uniform test (d) autocorrelation test, (e) multinomial chi-square test, (f) Poisson dispersion test, (g) Brown and Zhao chi-square test, and (h) big event triggering test. All of the tests can detect the simulations as nonrandom, though some tests perform slightly better than others. For example, the reduced sensitivity of the KS uniform test is due to the aftershock sequences being very localized in time, while the KS uniform test is more sensitive to long-term variations in seismicity rate. The autocorrelation and multinomial chi-square tests do not perform well at high magnitudes, indicating that these tests require more data than the others to detect nonrandom behavior.

4.2. Kolmogorov-Smirnov Tests

The KS methods compare the cumulative distribution function (CDF) determined from the data with the distribution expected for random event occurrence. The KS test computes a test statistic based on the largest absolute deviation between the two CDFs and then assigns a p value based on the test statistic and the number of observations in the data. For the KS exponential test [*Michael*, 2011], we look for a deviation between the expected and observed distributions of recurrence times, which is sensitive to short-term clustering in the data but not to long-term changes in the rate. Because the exponential distribution depends on the rate, which must be estimated from the data, we apply an appropriate correction based on Monte Carlo simulations [*Lilliefors*, 1969]. While both the variance test and the KS exponential test are based on the distribution of recurrence times, the individual events are combined in a different fashion in each test, and thus, the tests do not give identical results.

The KS uniform test [*Shearer and Stark*, 2012] differs from the KS exponential test in that it uses the occurrence times rather than the recurrence times. This difference is crucial, because the ordering of the events is important for the KS uniform test. Consider a hypothetical data set with a set of recurrence times that are exponentially distributed but ordered from shortest to longest. The KS exponential test would find the sequence consistent with random event occurrence, while the KS uniform test would not. This shows that the KS uniform test is more sensitive to long-term variations in the rate.

4.3. Autocorrelation Test

The final parameter-free test looks at the first lag of the autocorrelation of the recurrence times [*Michael*, 2011; *Parsons and Geist*, 2012]. If there are correlations between the recurrence times of consecutive events (as is

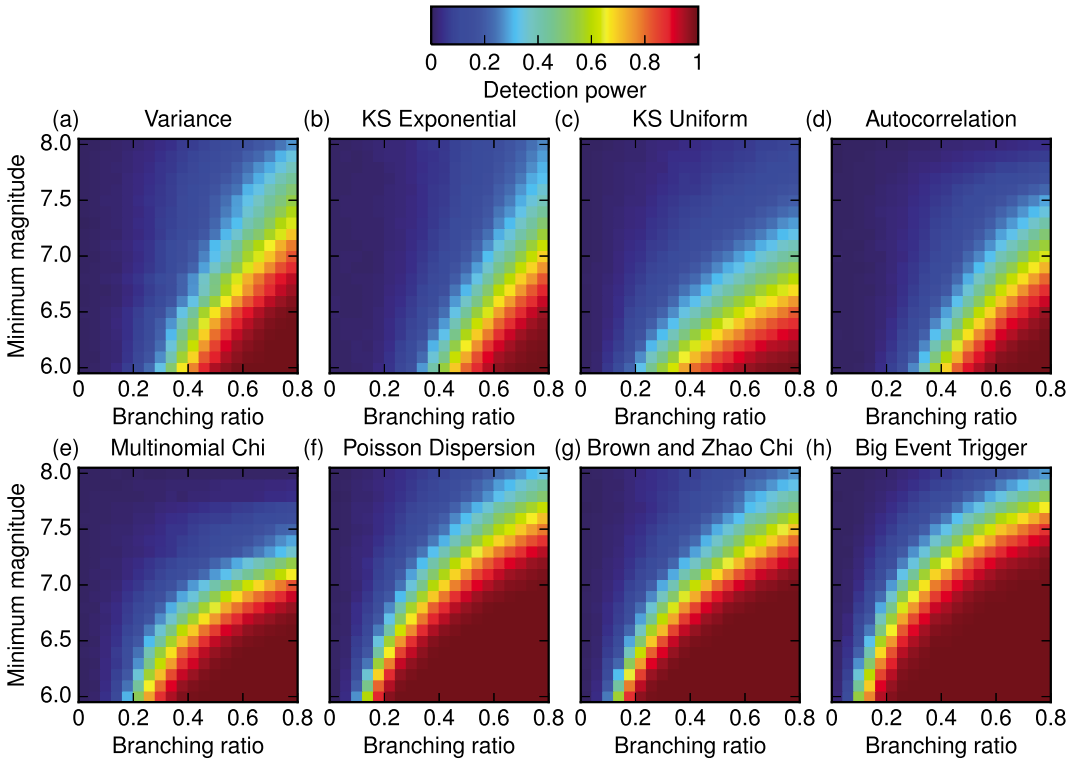

Figure 6. Same as Figure 5 but for the magnitude-dependent data sets. Because the mean rate COV is less than in the ETAS simulations, this clustering is more difficult to detect. The Poisson dispersion and big event triggering tests are the most successful for this type of clustering, as these tests look for variations in rate over longer time periods when compared to tests such as the variance, KS exponential, and autocorrelation, which use interevent times.

expected in an aftershock sequence), this test will identify the data as nonrandom. It is possible to include additional lags beyond the first in the test [*Parsons and Geist*, 2012], though for simplicity we only use the first lag in this study.

4.4. Multinomial Chi-Square Test

This test looks at the detailed distributions of the number of events occurring in a series of K time windows [*Gardner and Knopoff*, 1974; *Shearer and Stark*, 2012]. From the data, we use the observed number of events per time window λ and then determine the values of K^- and K^+. K^- is the smallest integer such that the expected number of windows with no more than K^- events is at least five, and K^+ is the largest integer such that the expected number of windows with at least K^+ events is at least five. Mathematically, this can be expressed as

$$K^- = \min\left\{ k : K \exp(-\lambda) \sum_{i=0}^{k} \frac{\lambda^i}{i!} \geq 5 \right\}, \tag{8}$$

$$K^+ = \max\left\{ k : K \left(1 - \exp(-\lambda) \sum_{i=0}^{k-1} \frac{\lambda^i}{i!} \right) \geq 5 \right\}. \tag{9}$$

From the values of K^- and K^+, the test calculates a test statistic based on the expected number of events per time bin E_k:

$$E_k = \begin{cases} K \exp(-\lambda) \sum_{i=0}^{K^-} \lambda^i / i!, & k = K^-, \\ K \exp(-\lambda)\, \lambda^k / k!, & K^- < k < K^+, \\ K \left(1 - \exp(-\lambda) \sum_{i=0}^{K^+-1} \lambda^i / i! \right), & k = K^+. \end{cases} \tag{10}$$

From the data, the test determines the observed number of time bins with fewer than K^- events X_{K^-}, the number of time bins with k events X_k, and the number of time bins with at least K^+ events X_{K^+}. These values are used to calculate the multinomial chi-square test statistic χ^2_M:

$$\chi^2_M = \sum_{k=K^-}^{K^+} \frac{(X_k - E_k)^2}{E_k}. \tag{11}$$

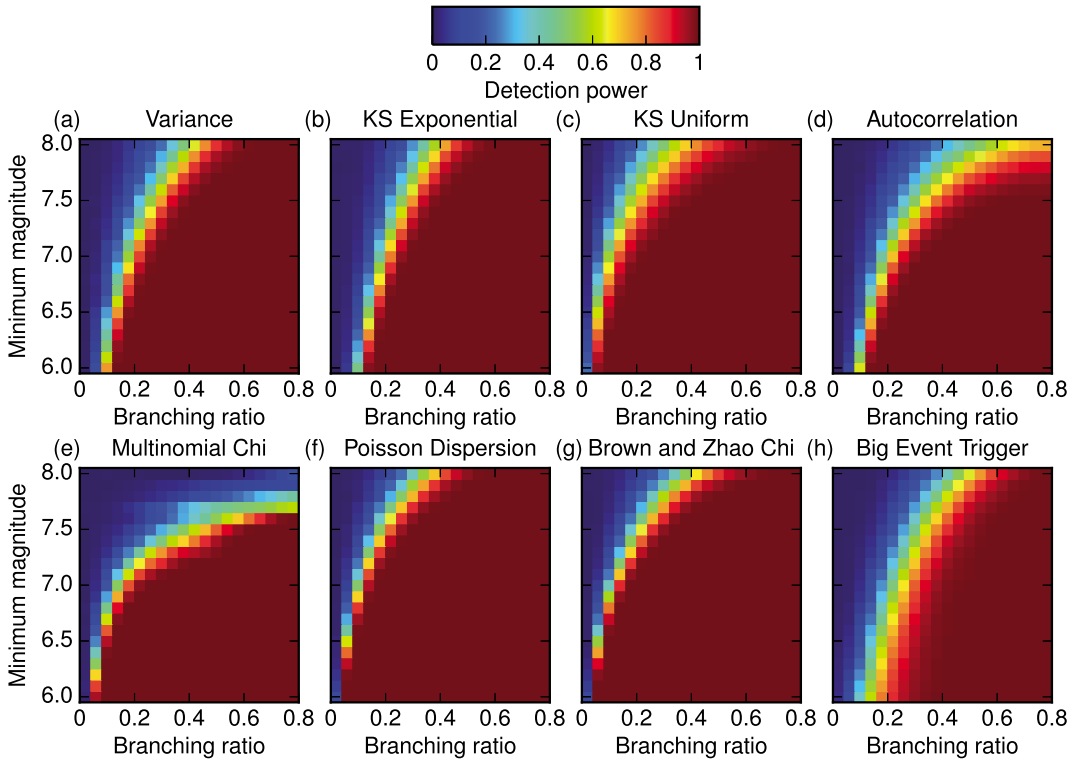

Figure 7. Same as Figure 5 but for the event clusters. All tests reliably detect clustering for these synthetic data sets, though the multinomial chi-square test does not do well at high magnitudes. Additionally, the big event triggering test requires more data than the other tests, due to the fact that the assumption that large events trigger small events does not hold for this particular type of clustering.

We choose 1 year for the time window, so the test examines if the distribution of the number of events in the time windows follows the expected distribution for random event occurrence. The original study by *Gardner and Knopoff* [1974] compared the distribution to a chi-square distribution, though here we follow *Shearer and Stark* [2012] and estimate the p values through Monte Carlo simulation using 10,000 random realizations. As with the variance test described above, for computational reasons we condition on the observed rate and duration, rather than the observed number of events, which are in agreement as long as the number of events is large.

4.5. Poisson Dispersion Test

The Poisson dispersion test (also referred to as a conditional chi-square test) divides the data set into discrete time bins and examines if the variance of the number of events per time bin is consistent with random event occurrence. Specifically, given K intervals, each containing N_k events, and the average number of events per window $\langle N_k \rangle$, the Poisson dispersion test calculates a test statistic

$$\chi_c^2 = \sum_{k=1}^{K} \frac{\left(N_k - \langle N_k \rangle\right)^2}{\langle N_k \rangle}. \tag{12}$$

We use 1 year for the time windows, which was used by both *Shearer and Stark* [2012] and *Daub et al.* [2012] on the global catalog. *Daub et al.* [2012] also looked at variable time windows and found that the results were not strongly dependent on the choice of time window. We estimate the p values by random simulation following the procedure of *Daub et al.* [2012] and condition on the observed rate and duration rather than the observed number of events as was done by *Shearer and Stark* [2012]. Both versions of the test when applied to the global earthquake catalog gave similar results. The test examines if there are an anomalous number of windows with a large number of events, which is indicative of clustering in the sequence of events being tested.

4.6. Brown and Zhao Chi-Square Test

This test is similar to the Poisson dispersion test but uses a slightly different test statistic that converges to a chi-square distribution with sufficient data [*Brown and Zhao*, 2002]. The test calculates the test statistic using

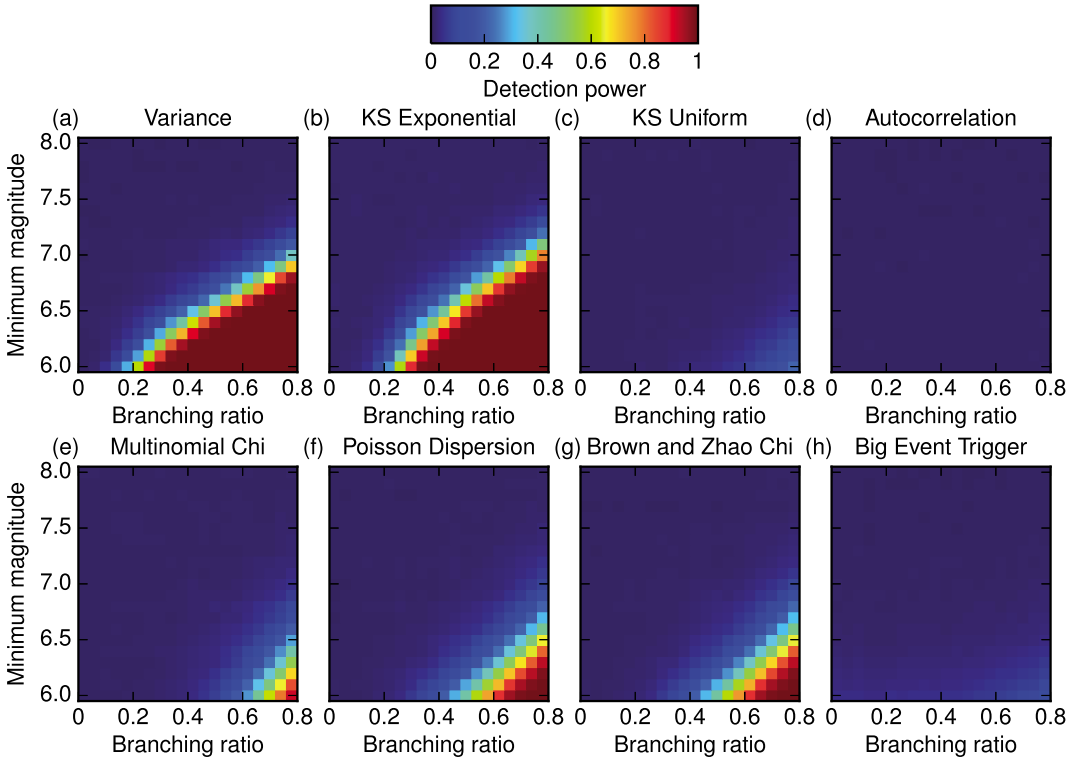

Figure 8. Same as Figure 5 but for the stochastic rate simulations. This type of clustering is the hardest to detect, and several tests do not find any nonrandom behavior in the synthetic data sets. The tests that are better able to discern that these sequences are nonrandom use the interevent time distribution (variance and KS exponential tests), though the autocorrelation test fails since successive recurrence times vary stochastically and thus are not correlated. The remaining tests perform poorly either due to no long-term variation in the rate or a lack of triggering of small events following large events in the case of the big event triggering test.

the number of observed events in a series of time windows and the observed seismicity rate. The test divides the data set into K windows, with N_k events in the kth window. The Brown and Zhao test statistic χ^2_{BZ} is then calculated from $Y_k = \sqrt{N_k + 3/8}$ and $\langle Y_k \rangle$ as

$$\chi^2_{BZ} = 4 \sum_{k=1}^{K} \left(Y_k - \langle Y_k \rangle \right)^2 . \tag{13}$$

We choose 1 year for the time window and estimate the p value by Monte Carlo simulation using 10,000 random realizations. As with the other tests, we condition on the rate and duration when using simulation to quantify the variability of the Brown and Zhao test statistic. This test is designed to be similar to the Poisson dispersion test, though we find in practice that the Poisson dispersion test tends to be more reliable at detecting the nonrandomness in our simulated data.

4.7. Big Event Triggering Test

The final test looks for a rate increase within a set of time windows following events above a chosen cutoff magnitude M_{big} [*Michael*, 2011]. We use a time window of 1 year and $M_{big} = 8.5$ in the test, though *Michael* [2011] examined the various parameter values more systematically. The test determines the number of events N_w that occur within the time windows following events with $M \geq M_{big}$ and then uses a binomial test to determine the probability of N_w events occurring in those windows if the true rate is always the observed seismicity rate (less the events with $M \geq M_{big}$, which are used to define the windows). If the number of small events that occur in the windows following the large events is greater than expected, then the test flags the data as nonrandom. Note that unlike the other tests, the big event triggering test makes an assumption about the specific mechanism of triggering (i.e., large events will tend to trigger small events). This may make the test more sensitive for detecting the nonrandom behavior that occurs in our simulations where large events trigger small events (ETAS, magnitude dependent) while reducing its sensitivity to clustering that does not follow the specific triggering model.

Table 3. Parameter Values for ETAS Models Varying the Magnitude-Frequency Distribution

Branching Ratio	Background Rate	$b = 0.8$ $M_{min} = 6$	$b = 1.2$ $M_{min} = 6$	$b = 1$ $M_{min} = 5$
0.02	98	0.0018	0.0035	0.0175
0.06	94	0.004	0.012	0.06
0.1	90	0.0065	0.021	0.1
0.14	86	0.0085	0.03	0.14
0.18	82	0.011	0.038	0.18
0.22	78	0.014	0.046	0.215
0.26	74	0.016	0.054	0.25
0.3	70	0.0185	0.063	0.29
0.34	66	0.02125	0.071	0.33
0.38	62	0.0235	0.079	0.365
0.42	58	0.026	0.088	0.405
0.46	54	0.0285	0.0965	0.44
0.5	50	0.031	0.1045	0.485
0.54	46	0.0335	0.113	0.525
0.58	42	0.03625	0.12	0.56
0.62	38	0.03875	0.128	0.6
0.66	34	0.041	0.138	0.64
0.7	30	0.0435	0.146	0.675
0.74	26	0.04625	0.154	0.715
0.78	22	0.04875	0.163	0.7575

5. Test Results

We test our synthetic data sets using the eight statistical tests described in section 4 at the 1% level. We threshold each sequence of events in magnitude at several levels, as is frequently done with the global earthquake record [*Michael*, 2011; *Shearer and Stark*, 2012; *Daub et al.*, 2012; *Ben-Naim et al.*, 2013], with magnitude levels ranging from $M \geq 6$ (the entire data set) to $M \geq 8$ with increments of 0.1 magnitude units. This allows us to determine the likelihood of detection for data sets with between 10,000 and 100 events on average, depending on the magnitude level chosen.

Each type of synthetic data set has 20 versions with different clustering strengths, and each of the 20 versions is simulated 10,000 times. For each set of magnitude level and clustering strength, we determine the detection power, in other words the probability that the test identifies the synthetic data set as nonrandom. If the detection power is nearly unity, then the test reliably identifies the nonrandom aspects of the simulated data, while if the detection power is close to zero, then the test cannot distinguish between the simulated events and a random sequence of events.

We summarize the results of the statistical tests in Figures 5–8. Each subplot of the figures shows the magnitude level on the vertical scale and branching ratio on the horizontal scale, with the color scale indicating detection power. Figures 5–8 (a) show the detection power for the variance test, (b) illustrate the detection power for the KS exponential test, (c) depict the detection power for the KS uniform test, (d) show the results for the autocorrelation test, (e) illustrate the detection power of the multinomial chi-square test, (f) depict the results for the Poisson dispersion test, (g) show the Brown and Zhao chi-square test, and (h) illustrate the detection power of the big event triggering test. The results for the ETAS simulations are shown in Figure 5, the results for the magnitude-dependent simulations are illustrated in Figure 6, the results for the event clusters can be found in Figure 7, and the results for the stochastic rate simulations are shown in Figure 8.

As expected, all tests show improved detection power as the strength of the clustering increases, and performance decreases as the magnitude threshold increases (and the number of events decreases). We also find the expected result that some tests are better suited for detecting certain types of nonrandom behavior than others. For example, the KS uniform, autocorrelation, and multinomial do not perform as well as the other tests

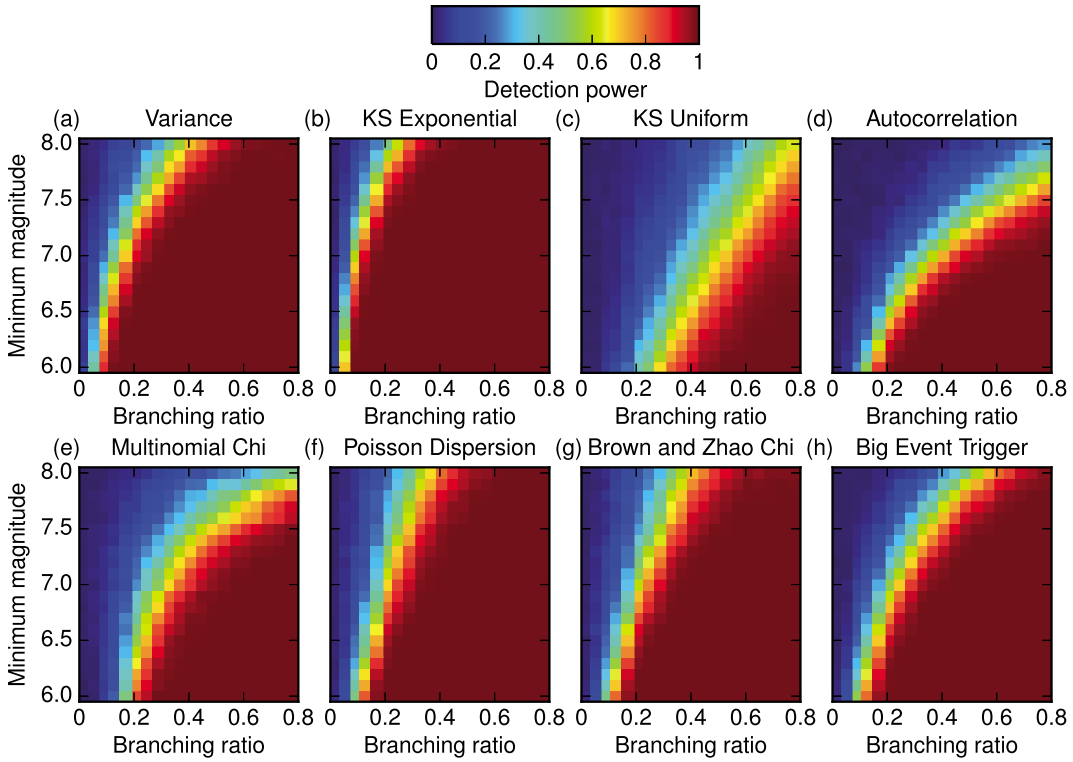

Figure 9. Detection power at the 1% level as a function of branching ratio and minimum magnitude for three statistical tests applied to the ETAS simulations with $b = 0.8$. The tests are better able to distinguish the simulation data as nonrandom for high magnitudes when compared to the $b = 1$ results, due to the increased number of high-magnitude events in the $b = 0.8$ simulations.

on the ETAS simulations. The KS uniform test is best at detecting long-term variations in the seismicity rate, while the aftershock sequences that characterize ETAS simulations are localized in time and do not introduce a change in the seismicity rate over long time periods. Conversely, the KS uniform test performs better than the KS exponential test for the magnitude-dependent simulations. This is because the magnitude-dependent rate change extends over longer periods of time, which is more difficult to detect when using the interevent time distribution.

The nonrandomness of the event clusters is easily detected by all of the tests, though the multinomial chi-square test does not perform well at high magnitudes, as it looks at the detailed distribution of events per time bin, details that cannot be discerned for small numbers of events. All tests are successful because this type of clustering introduces both long-term variations in the rate and many extra short recurrence times during the clusters. Interestingly, the big event triggering test is still able to flag the event clusters data sets as nonrandom, despite making the incorrect assumption that the seismicity rate increases following large events. The test can identify these data sets as nonrandom because it preferentially samples the elevated rate of smaller events during the two clusters, due to the fact that the events with M_{big} are also more likely to occur during the clusters. However, it does require more data than many of the other tests, so it does suffer to some degree for its incorrect assumption. On the other hand, the stochastic rate data sets are the most difficult to identify as nonrandom, as the rate changes are variable, being neither localized in time (like the ETAS simulations) nor extended in time (like the magnitude-dependent and event clusters data sets). Only the tests that use the distribution of the recurrence times (the variance and KS exponential tests) have much success with this type of clustering. The other tests fail due to a lack of a long-term rate change (KS uniform), no correlation between successive recurrence times (autocorrelation), lack of short-term clustering over time scales of ~1 year (multinomial, Poisson dispersion, and Brown and Zhao tests), and no relationship between large events and seismicity rate increases (big event triggering).

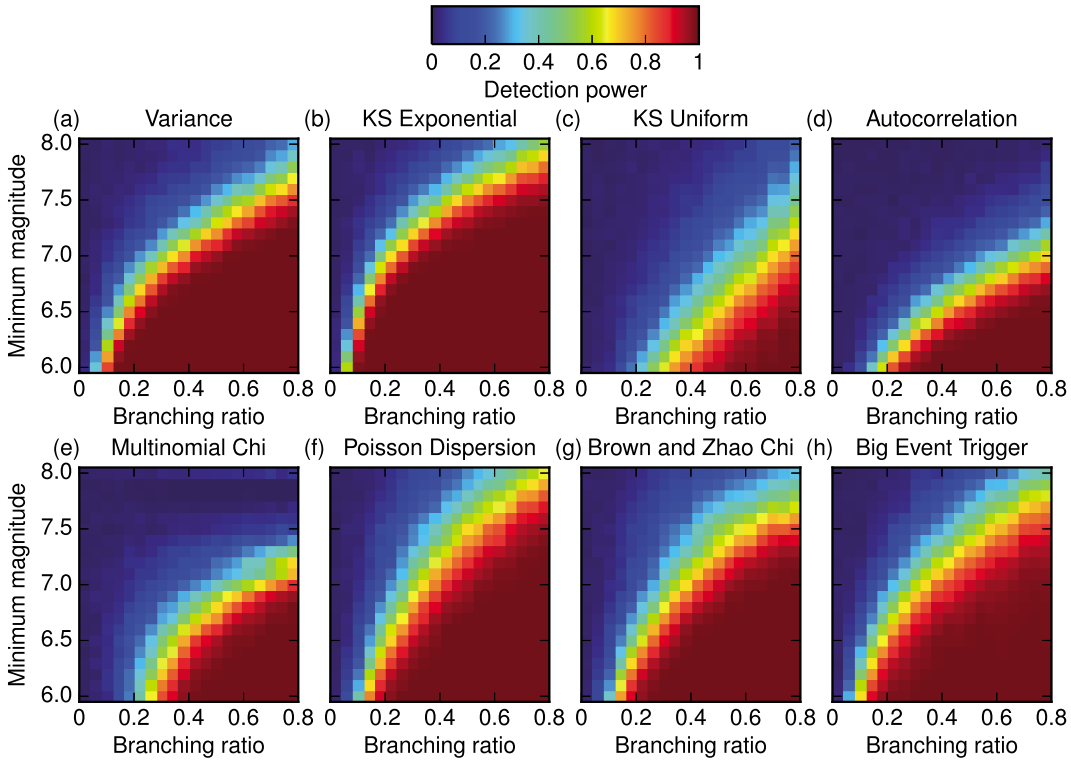

Figure 10. Detection power at the 1% level as a function of branching ratio and minimum magnitude for three statistical tests applied to the ETAS simulations with $b = 1.2$. The tests do not perform as well as the tests on the $b = 1$ results at high magnitudes, as the larger b value leads to relatively fewer high-magnitude events.

5.1. Sensitivity to Magnitude-Frequency Distribution

Because of the differences in the magnitude-frequency distribution between the PAGER and GEM catalogs, we perform additional tests using ETAS simulations that are generated using varying magnitude-frequency distributions. In particular, we run two additional sets of simulations maintaining $M_{\text{min}} = 6$ but changing the b value to 0.8 or 1.2 and one additional simulation with $b = 1$ but $M_{\text{min}} = 5$. In the case of the variable b values, the relative number of small versus large events changes, so we can examine if the magnitude distribution within an aftershock sequences affects the ability of the various statistical tests to identify the nonrandom character of that aftershock sequence. Similarly, the $M_{\text{min}} = 5$ simulations test whether or not we bias our results by not simulating the smaller events that exist in natural seismicity sequences below the detection threshold in a catalog. The parameter values for these three additional ETAS models are shown in Table 3 and are chosen to maintain the overall rate of 100 events/year in our simulated data sets. Note that by changing the b value, the number of events at higher magnitudes is different (the number decreases for $b = 1.2$ relative to the $b = 1$ simulations, and the number increases for $b = 0.8$ relative to the $b = 1$ simulations), while the overall number of events at all magnitude levels remains the same for the $M_{\text{min}} = 5$ simulations.

The results of this analysis are shown in Figures 9–11. Each figure shows the same set of plots as described in Figure 5 but with a different magnitude-frequency distribution: Figure 9 shows the results for $b = 0.8$ and $M_{\text{min}} = 6$, Figure 10 illustrates the detection power for the various statistical tests for $b = 1.2$ and $M_{\text{min}} = 6$, and Figure 11 shows the detection capabilities of the tests for the ETAS simulations with $b = 1$ and $M_{\text{min}} = 5$. The detection power changes somewhat for the varying b values, but the differences can be attributed entirely to changes in the number of events. When $b = 0.8$, there are more large-magnitude events, and thus, we find that the statistical tests have greater detection power when compared to the $b = 1$ results, and when $b = 1.2$, the reverse is true. For simulations with $M_{\text{min}} = 5$, the results are essentially the same as for $M_{\text{min}} = 6$. These additional simulations suggest that our results should be broadly applicable to other earthquake catalogs with different magnitude-frequency distributions, as it is primarily the number of events that controls what clustering can be detected.

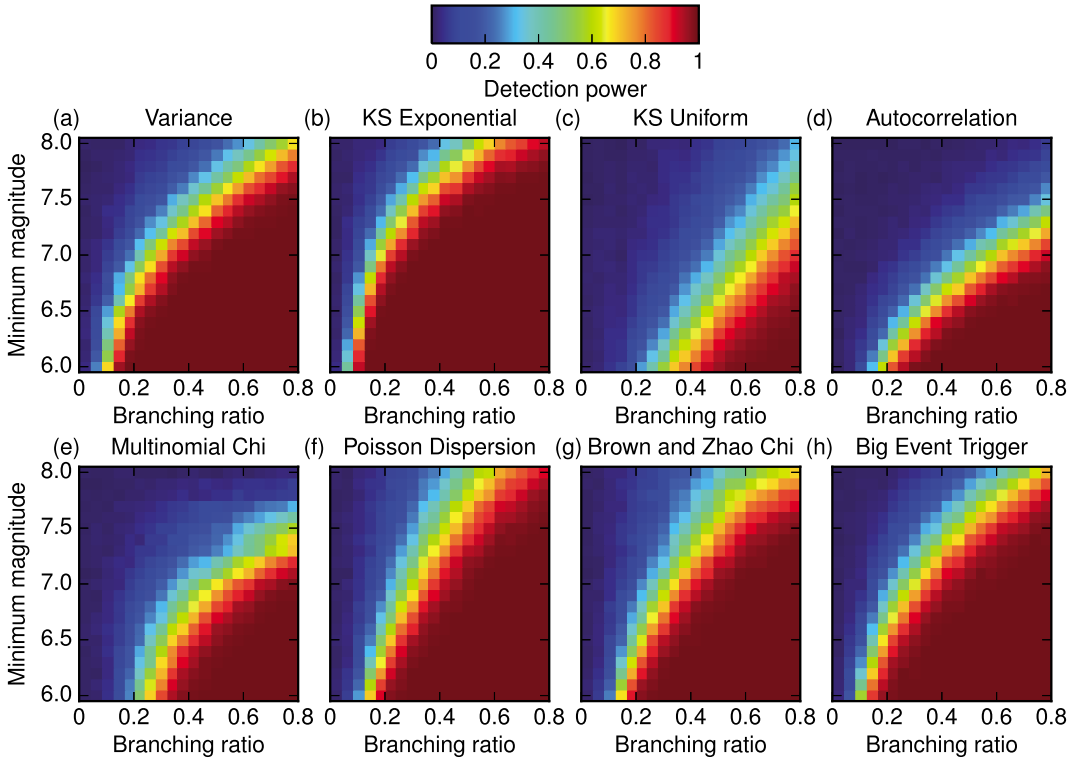

Figure 11. Detection power at the 1% level as a function of branching ratio and minimum magnitude for eight statistical tests applied to the ETAS simulations with $M_{min} = 5$. The results do not show much difference from those in Figure 5, indicating that the detection power of the statistical tests is not changed by the fact that real earthquake catalogs are missing smaller-magnitude events.

6. Analysis of *p* Values

The detection power analysis of the various statistical tests applied to the synthetic data sets provides information on the clustering strengths that are detectable in the global earthquake data set. However, this analysis requires a choice of the *p* value that is considered to be statistically significant, an issue on which there is often some debate. To avoid the problem of choosing a significance level, we instead analyze the *p* values that do not give a statistically significant result to see how much information the *p* values contain about the clustering level. If *p* values turn out to be predictive of clustering strength, then instead of picking a significance level, one could simply test the data and then infer a likely range of clustering strengths based on the *p* value. This approach can be thought of as applying a Bayesian framework to the problem, with the clustered data sets producing the prior distribution. Through our analysis, we can use our simulation results to find the posterior distribution of the clustering strength of the earthquake record conditioned on the *p* value observed by applying the statistical test to the global earthquake record.

To implement this framework, we analyze the *p* values that do not constitute a positive test (i.e., $p > 0.01$) for several different magnitude levels, tests, and clustering types. We take all synthetic realizations for which $p > 0.01$ and then bin the *p* values specific to each catalog type in five different bins which are logarithmically spaced, centered at $p = 0.015, 0.038, 0.095, 0.24$, and 0.602. Each bin contains simulations that are potentially from any clustering level. We can then determine the cumulative distribution (CDF) as a function of clustering strength for each set of *p* values. The CDF for each bin shows the likelihood that the *p* values in that particular bin are drawn from any of the clustering strengths above that particular branching ratio and thus can be used to directly infer upper limits on the branching ratio at particular confidence levels. If the CDF falls rapidly from unity to zero, then there is a strong correlation between *p* value and clustering level and *p* values have predictive power. On the other hand, if the decrease in the CDF from unity to zero is more gradual, then there is a weaker correlation between *p* value and clustering level and little can be learned from the *p* value.

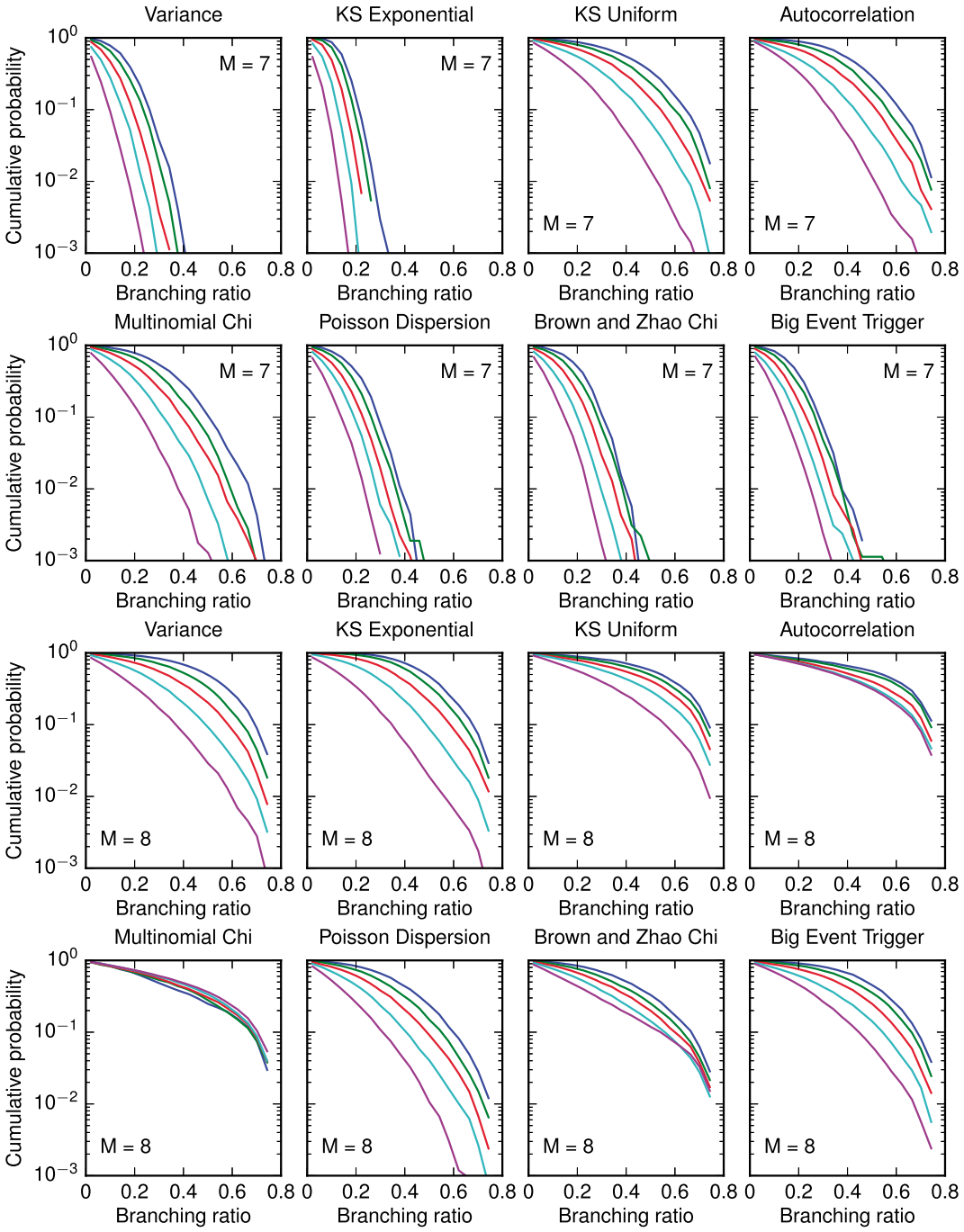

Figure 12. Cumulative distribution functions for ETAS simulation p values that were not identified as nonrandom by the statistical tests (i.e., $p > 0.01$). The statistical test used in each set of CDF functions is indicated above the plot. In each case, the plots show the CDF as a function of branching ratio for five bins of p values, centered at the values of 0.015, 0.038, 0.095, 0.240, and 0.602 (p values increase from top to bottom on each set of curves). (first and second rows) Results for $M = 7$ and (third and fourth rows) results for $M = 8$. Tests that are more likely to detect a synthetic data set as nonrandom also tend to have p values that are more predictive of clustering level than tests with lower detection power. The CDF can be easily used to place an upper bound on the level of clustering that is consistent with the p value observed for the global earthquake record—for instance, the KS exponential test applied to the PAGER/PDE catalog at $M = 7$ gives $p = 0.17$ [*Michael*, 2011], which is the second curve from the bottom, and we can thus constrain the branching ratio for aftershock-like behavior to be below 0.2 at the 99th percentile. Bounds on the branching ratio at $M = 8$ are higher due to the small number of events in the simulated data.

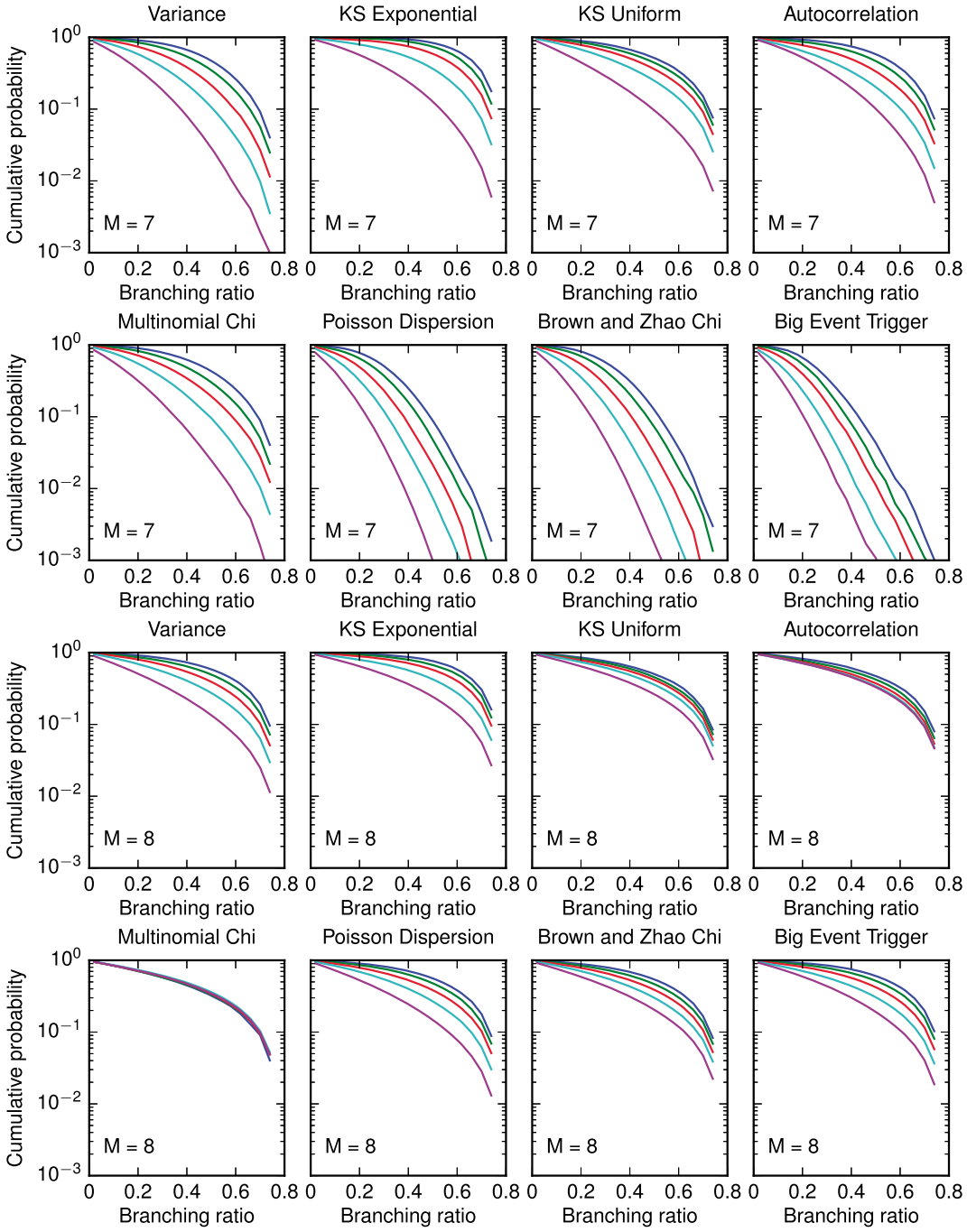

Figure 13. Same as Figure 12 but for the magnitude-dependent simulations. Because this clustering is harder to detect, the CDFs do not fall off as sharply with increasing branching ratio when compared to the ETAS simulations. In several cases, the *p* values have very little predictive power of the level of clustering in the data, as the CDF curves fall off slowly with branching ratio, and the curves for different *p* values are nearly identical to one another.

Example CDF functions calculated using this analysis are shown in Figure 12. The plots illustrate the CDF as a function of branching ratio for the five bins of *p* values for the ETAS simulations. The *p* value of each bin decreases from the bottom curve to the top curve. Each plot shows the results for one particular statistical test, and the top set of plots show the CDFs for $M = 7$ and the bottom set of plots are for $M = 8$.

The CDF plots illustrate several aspects of the test results. For the tests that perform well on the ETAS simulations, the CDF falls off fairly quickly as the branching ratio increases for all *p* values. This indicates that for

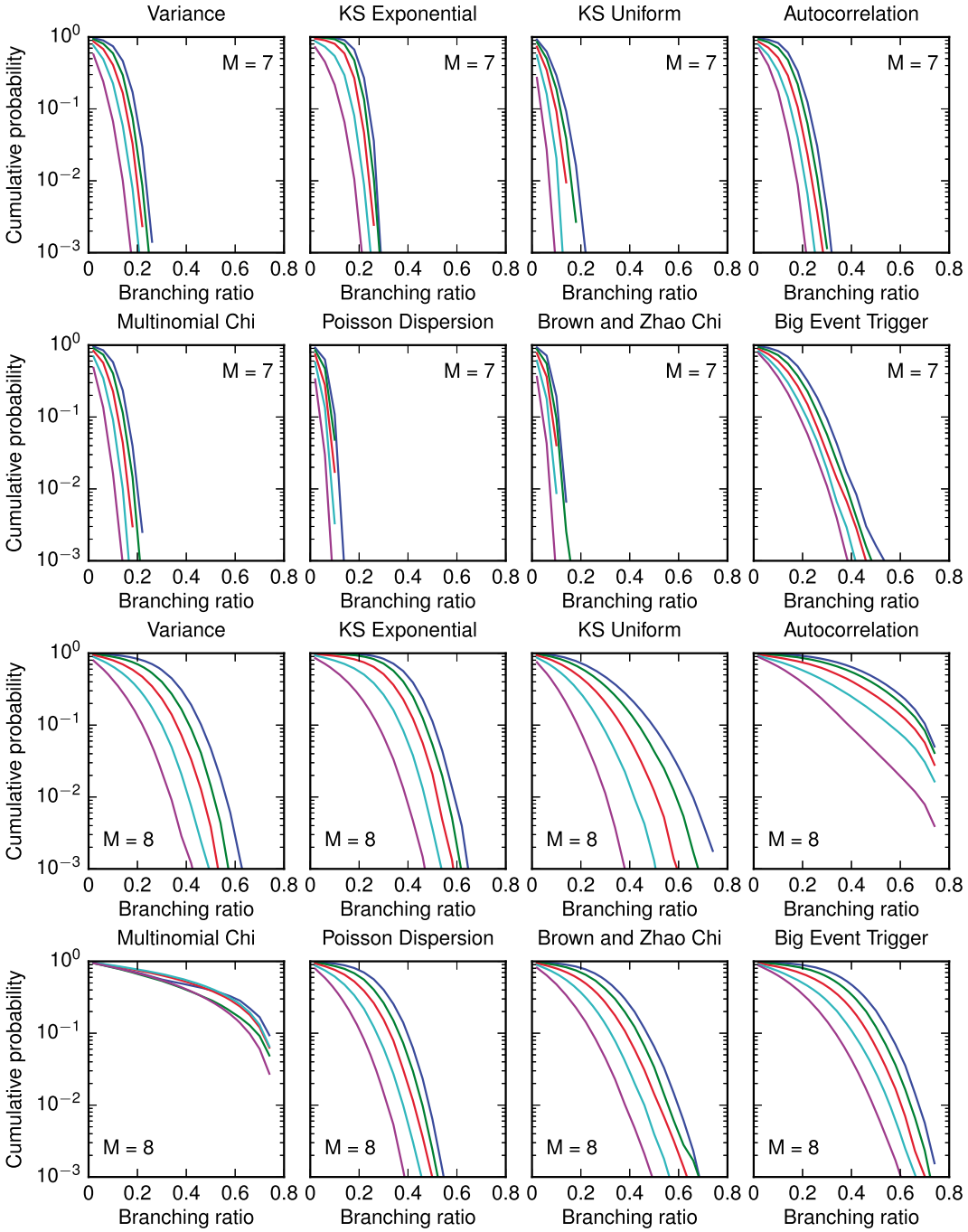

Figure 14. Same as Figure 12 but for the event clusters. With the exception of the multinomial chi-square test at high magnitudes, nearly all tests are able to provide constraints regarding the clustering level, though tests that perform better are better able to constrain an upper bound on the level of clustering in the synthetic data set. Several of the tests at $M = 7$ suggest that the largest clustering strength consistent with the global earthquake record is 0.1 at the 99th percentile, while at $M = 8$ the upper bound on the branching ratio is below 0.35 at the 99th percentile using the CDFs for the Poisson dispersion or KS uniform tests.

many statistical tests, the p value is well correlated with the level of clustering even if the p value is not low enough to constitute a statistically significant result. At $M \geq 8$ the CDFs decrease more gradually. This is due to the reduced number of events at this magnitude level, and so these results are less useful for constraining the clustering level.

These trends are further illustrated for data sets where the clustering is more difficult to detect. Figure 13 shows the same CDF plots as in Figure 12 for the magnitude-dependent synthetic data sets. For the magnitude-dependent simulations, the big event triggering test performs best, though a number of other tests also perform well, and this is reflected in the dropoff of the CDF curves. For the magnitude-dependent simulations, the tests exhibiting reduced detection power have CDF curves that fall off gradually and show little difference across the p value bins. This shows that for tests where detection of the nonrandom character of the synthetic data is less likely, there is also little correlation between p values and the level of clustering in the data set. For the stochastic rate simulations (not shown), nearly all of the CDF curves provide little information on clustering strength.

Figure 14 shows the same sets of CDFs for the event clusters. Here all of the tests perform well and produce CDFs that fall off rapidly with branching ratio. There are differences in precisely how fast the curves decrease across different tests, but, in general, the p values of each test provide information on the clustering strength. At $M = 7$, the steep drop in the CDFs with branching ratio indicates that we can use our results to place quantitative upper bounds on the branching ratio. At $M = 8$, the CDFs decrease more slowly but can still provide upper bounds on the clustering strength in the global catalog.

The general trends found in our results help us understand what we can learn from the p values in the global earthquake catalog. Because all of these tests have been previously applied to the earthquake data since 1900, where the level or type of clustering is unknown, we can use these results to bound the clustering level that is likely to be present in the data, which we discuss in the following section.

7. Discussion

The simulations used in this study show that statistical tests cannot always detect that an earthquake data set is nonrandom. There are two effects responsible for this. First, tests need a certain amount of data in order to distinguish nonrandom event occurrence from the expected fluctuations in a random process. While this result is unsurprising, our results confirm that this is the case for the earthquake catalog at high-magnitude levels, and we have quantified how this depends on the strength of the clustering. Second, we find that not all tests perform equally well on a given type of clustering and that if the wrong test is performed, the test exhibits slower improvement as more data are added than a test that is better suited to identifying the given type of clustering. This suggests that performing the correct test is just as important as having enough data, as it makes a given amount of data more useful.

This question has been previously studied by *Dimer de Oliveira* [2012], though only for higher-magnitude levels and one type of clustering (similar to our event clusters simulations). The clustered data sets of the Dimer de Oliveira study varied the background rate from 0.1 events/year to 0.5 events/year, with a tenfold increase in the seismicity rate during a number of 15 year clusters during a 110 year event sequence, with the number of clusters ranging from 1 to 5. This best corresponds to our event clusters simulations at $M = 8$ at branching ratios of 0.62 or 0.66, with a background rate of 0.38 or 0.34 events/year, respectively, and two 10 year clusters where the event rate is 3.4 or 3.6 events/year, respectively. Dimer de Oliveira found that for four events/decade and two clusters, the detection power at the 5% level was about 40%. In our case, we found that we could almost always detect that these simulated data sets are clustered at the 1% level, though at slightly lower clustering levels our detection power diminishes quickly. These differences in detection power may be due to differences in the exact number of events or the fact that we generate our data using the event rate at $M = 6$ rather than $M = 8$, resulting in a different amount of variability in the number of large-magnitude events across different realizations.

Our results have direct relevance for studies that compare the global earthquake record with a process that is random in time. The results of such tests tend to find that the global catalog does not deviate from a random process at p values that are considered to be significant, and thus, our results can bound the range of clustering that could be in the data yet not be detectable.

We can use the CDFs constructed for different p values (i.e., Figures 12 and 14) to provide quantitative upper bounds on the branching ratio given the results of tests applied to the earthquake catalog. For instance, *Michael* [2011] reports $p = 0.17$ for the earthquake record for $M \geq 7$ with aftershocks removed. That value falls within our $p = 0.240$ bin (though at the low end of the bin). Based on Figure 12, this suggests that the percentage of events that are aftershocks is below 20% based on the 99th percentile bound, as the CDF drops below 10^{-2} at a branching ratio of 0.2 for this particular test. Other statistical tests give similar upper bounds when applied to the PAGER/PDE catalog, so this appears to be a robust upper limit based on several statistical tests.

We can make a similar estimate for event clusters simulations using the results in Figure 14. The best performing test on this particular type of clustering was the Poisson dispersion test. *Shearer and Stark* [2012] were more conservative in their removal of aftershocks, leaving only 509 independent events in their catalog for $M \geq 7$, which corresponds to approximately $M \geq 7.3$ in our simulations. These estimates constrain the branching ratio to be below 0.1 at the 99th percentile if magnitude 7 events occur in two large clusters using the Poisson dispersion test. In terms of seismicity rate during the clusters, a branching ratio of 0.1 at $M = 7$ indicates a ∼50% increase in seismicity rate (from 9 events/year to 14 events/year during the clusters). We find that clustering stronger than this is unlikely to give us the observed p values. Other tests give similar upper bounds on the maximum branching ratio that is consistent with the PAGER/PDE data.

At higher-magnitude levels above $M \geq 8$, we cannot constrain the branching ratio based on the p values of statistical tests to the same degree as $M = 7$. While synthetic data sets that are more clustered tend to have smaller p values, the correlation is not strong enough to rule out higher branching ratios with high confidence. The study of *Michael* [2011] on the $M \geq 8$ earthquake data reported $p = 0.61$ for the KS exponential test. Our results for the ETAS models for that range of p values suggest that the branching ratio of the global catalog is below 0.6 at the 99th percentile, a much broader range than we infer for the $M \geq 7$ data. Similarly, *Shearer and Stark* [2012] found $p = 0.898$ for the Poisson dispersion test at $M \geq 8$, which suggests a branching ratio below 0.6 at the 99th percentile for the ETAS simulations but an upper bound of 0.35 at the 99th percentile for the event clusters data sets. The KS uniform test applied to the event clusters simulations provides a similar estimate of these bounds. In terms of seismicity rate increases for $M = 8$, a branching ratio of 0.35 corresponds to an increase from 0.66 events/year to 2.3 events/year during the clusters. Clustering strengths higher than this would be detectable given the amount of data in the PAGER/PDE catalog, but lower values produce sequences of events that are not reliably distinguishable from random event occurrence.

We note that each of the clustering studies discussed above used a different method to remove aftershocks, while our simulations are based on the assumption that there are no local aftershocks in the synthetic data. Because aftershock removal is somewhat subjective, there is a good chance that some background events are removed in the process (see *Luen and Stark* [2012] for a discussion of some of the issues raised by the aftershock removal process). Our simulations do not consider the spatial distribution of earthquakes and thus do not exhibit this artifact of declustering. Further studies, for example, using ETAS simulations with spatial kernels, are necessary to better quantify the impact of this effect on the results of statistical tests.

One caveat of our analysis is that our synthetic data sets are only four of any number of possible ways that events can cluster in time. We have used various observations and models for clustering to guide our development of our simulations, but this is only a partial, limited set of considerations. Earthquake data may also contain a combination of multiple types of clustering, which may affect the ability of statistical tests to identify nonrandom behavior, and statistical tests cannot reveal the underlying mechanism of clustering, only the likelihood that it is present. Because the rules for aftershock production are based on more robust observations than those for the other types of clustering, the ETAS simulations may be more relevant to seismic hazard estimates. This is especially true since the effect of aftershocks is more localized in time than the other types of clustering, and aftershock forecasts are one of the few types of short-term seismic hazard estimates that can be made with confidence [*Jordan et al.*, 2011]. Thus, such analysis can be extended using the upper bounds on global aftershock production to estimate the effect such global aftershocks would have on seismic hazard. Such an estimate would only serve as an upper bound but could inform us whether such effects might play a role in hazard estimates.

While our focus here is on the global earthquake catalog, our methods can also be applied to regional catalogs at lower magnitude levels. Previous work has mostly focused on earthquake catalogs in Southern California [*Gardner and Knopoff*, 1974; *Luen and Stark*, 2012], Japan [*Zhuang et al.*, 2004], and Taiwan [*Wang et al.*, 2014],

but as seismic networks continue to grow worldwide, more regional catalogs will become available. For example, our CDF-based methods could be used to provide an assessment of the goodness of fit between ETAS models and regional catalogs and help constrain regional aftershock parameters [*Ogata*, 1992, 1998]. Knowledge of how to analyze such catalogs for seismicity patterns and develop models for earthquake interaction requires quantification of clustering that is consistent with the catalog. The tools outlined here provide a means of doing that and can help researchers bound the strength of earthquake interactions in a variety of tectonic settings.

Acknowledgments
We thank Andrew Michael and an anonymous reviewer for their constructive reviews. This research was supported by DOE grant DE-AC52-06NA25396 and institutional (LDRD) funding at Los Alamos. Figures were generated using the Python plotting library Matplotlib [*Hunter*, 2007].

References

Aki, K. (1965), Maximum likelihood estimate of b in the formula $\log N = a - bM$ and its confidence limits, *Bull. Earthq. Res. Inst. Univ. Tokyo*, *43*, 237–239.
Allen, T. I., K. Marano, P. S. Earle, and D. J. Wald (2009), PAGER-CAT: A composite earthquake catalog for calibrating global fatality models, *Seismol. Res. Lett.*, *80*, 50–56.
Ben-Naim, E., E. G. Daub, and P. A. Johnson (2013), Recurrence statistics of great earthquakes, *Geophys. Res. Lett.*, *40*, 3021–3025, doi:10.1002/grl.50605.
Brown, L. D., and L. H. Zhao (2002), A test for the Poisson distribution, Sankhyā: Indian, *J. Stat.*, *64*, 611–625.
Bufe, C. G., and D. M. Perkins (2005), Evidence for a global seismic-moment release sequence, *Bull. Seismol. Soc. Am.*, *95*, 833–843.
Daub, E. G., E. Ben-Naim, R. A. Guyer, and P. A. Johnson (2012), Are megaquakes clustered?, *Geophys. Res. Lett.*, *39*, L06308, doi:10.1029/2012GL051465.
Dimer de Oliveira, F. (2012), Can we trust earthquake cluster detection tests?, *Geophys. Res. Lett.*, *39*, L17305, doi:10.1029/2012GL052130.
Felzer, K. R., R. E. Abercrombie, and G. Ekstrom (2004), A common origin for aftershocks, foreshocks, and multiplets, *Bull. Seismol. Soc. Am.*, *94*, 88–98.
Felzer, K. R., and E. E. Brodsky (2006), Decay of aftershock density with distance indicates triggering by dynamic stress, *Nature*, *441*, 735–738.
Freed, A. M. (2005), Earthquake triggering by static dynamic, and postseismic stress transfer, *Annu. Rev. Earth Planet. Sci.*, *33*, 335–367, doi:10.1146/annurev.earth.33.092203.122505.
Gardner, J. K., and L. Knopoff (1974), Is the sequence of earthquakes in Southern California, with aftershocks removed, Poissonian?, *Bull. Seismol. Soc. Am.*, *64*, 1363–1367.
Gomberg, J., P. Bodin, K. Larson, and H. Dragert (2004), Earthquake nucleation by transient deformations caused by the M = 7.9 Denali, Alaska, earthquake, *Nature*, *427*, 621–624.
Gomberg, J., P. Bodin, and P. A. Reasenberg (2003), Observing earthquakes triggered in the near field by dynamic deformations, *Bull. Seismol. Soc. Am.*, *93*, 118–138.
Guo, Z., and Y. Ogata (1997), Statistical relations between the parameters of aftershocks in time, space, and magnitude, *J. Geophys. Res.*, *102*(B2), 2857–2873, doi:10.1029/96JB02946.
Gutenberg, B., and C. F. Richter (1954), *Seismicity of the Earth and Associated Phenomena*, 2nd ed., Princeton Univ. Press, Princeton.
Helmstetter, A., and D. Sornette (2002), Subcritical and supercritical regimes in epidemic models of earthquake aftershocks, *J. Geophys. Res.*, *107*(B10), 2237, doi:10.1029/2001JB001580.
Helmstetter, A., and D. Sornette (2003), Båth's law derived from the Gutenberg-Richter law and from aftershock properties, *Geophys. Res. Lett*, *30*(20), 2069, doi:10.1029/2003GL018186.
Helmstetter, A., Y. Y. Kagan, and D. D. Jackson (2005), Importance of small earthquakes for stress transfers and earthquake triggering, *J. Geophys. Res.*, *110*, B05S08, doi:10.1029/2004JB003286.
Hill, D. P., et al. (1993), Remote seismicity triggered by the M7.5 Landers, California earthquake of June 28, 1992, *Science*, *260*, 1617–1623.
Hunter, J. D. (2007), Matplotlib: A 2D graphics environment, *Comput. Sci. Eng.*, *9*(3), 90–95.
Jordan, T., Y. Chen, P. Gasparini, R. Madariaga, I. Main, W. Marzocchi, G. Papadopoulos, G. Sobolev, K. Yamaoka, and J. Zschau (2011), Operational earthquake forecasting: State of knowledge and guidelines for utilization, *Ann. Geophys.*, *54*(4), 316–391, doi:10.4401/ag-5350.
Lay, T., and T. C. Wallace (1995), *Modern Global Seismology*, Academic, San Diego, Calif.
Lilliefors, H. W. (1969), On the Kolmogorov-Smirnov test for the exponential distribution with mean unknown, *J. Am. Stat. Assoc.*, *64*, 387–389, doi:10.2307/2283748.
Luen, B., and P. B. Stark (2012), Poisson tests of declustered catalogues, *Geophys. J. Int.*, *189*, 691–700, doi:10.1111/j.1365-246X.2012.05400.x.
Michael, A. J. (2011), Random variability explains apparent global clustering of large earthquakes, *Geophys. Res. Lett.*, *38*, L21301, doi:10.1029/2011GL049443.
Michael, A. J. (2014), How complete is the ISC-GEM global earthquake catalog?, *Bull. Seismol. Soc. Am.*, *104*(4), 1829–1837, doi:10.1785/0120130227.
Ogata, Y. (1992), Detection of precursory relative quiescence before great earthquakes through a statistical model, *J. Geophys. Res.*, *97*(B13), 19,845–19,871, doi:10.1029/92JB00708.
Ogata, Y. (1998), Space-time point-process models for earthquake occurrences, *Ann. Inst. Stat. Math.*, *50*(2), 379–402.
Omori, F. (1895), On the aftershocks of earthquakes, *J. Coll. Sci., Imp. Univ. Tokyo*, *7*, 111–200.
Parsons, T., and E. L. Geist (2012), Were global $M \geq 8.3$ earthquake time intervals random between 1900 and 2011?, *Bull. Seismol. Soc. Am.*, *102*(4), 1583–1592, doi:10.1785/0120110282.
Parsons, T., and E. L. Geist (2014), The 2010–2014.3 global earthquake rate increase, *Geophys. Res. Lett.*, *41*, 4479–4485, doi:10.1002/2014GL060513.
Parsons, T., and A. A. Velasco (2011), Absence of remotely triggered large earthquakes beyond the mainshock region, *Nat. Geosci.*, *4*, 312–316.
Pollitz, F. F., R. S. Stein, V. Sevilgen, and R. Bürgmann (2012), The 11 April 2012 east Indian Ocean earthquake triggered large aftershocks worldwide, *Nature*, *490*, 250–253, doi:10.1038/nature11504.
Pollitz, F. F., R. Bürgmann, R. S. Stein, and V. Sevilgen (2014), The profound reach of the 11 April 2012 M 8.6 Indian Ocean earthquake: Short-term global triggering followed by a longer-term global shadow, *Bull. Seismol. Soc. Am.*, *104*(2), 972–984, doi:10.1785/0120130078.
Rhoades, D. A., and F. F. Evison (2004), Long-range earthquake forecasting with every earthquake a precursor according to scale, *Pure Appl. Geophys.*, *161*(1), 47–72, doi:10.1007/s00024-003-2434-9.

Shearer, P. M., and P. B. Stark (2012), The global risk of big earthquakes has not recently increased, *Proc. Natl. Acad. Sci.*, *109*(3), 717–721.
Sornette, A., and D. Sornette (1999), Renormalization of earthquake aftershocks, *Geophys. Res. Lett.*, *26*, 1981–1984, doi:10.1029/1999GL900394.
Storchak, D. A., D. Di Giacomo, I. Bondär, E. R. Engdahl, J. Harris, W. H. K. Lee, A. Villaseñor, and P. Bormann (2013), Public release of the ISC-GEM global instrumental earthquake catalogue (1900–2009), *Seismol. Res. Lett.*, *84*(5), 810–815, doi:10.1785/0220130034.
Utsu, T., Y. Ogata, and R. S. Matsu'ura (1995), The centenary of the Omori formula for a decay law of aftershock activity, *J. Phys. Earth*, *43*, 1–33.
van der Elst, N. J., and E. E. Brodsky (2010), Connecting near-field and far-field earthquake triggering to dynamic strain, *J. Geophys. Res.*, *115*, B07311, doi:10.1029/2009JB006681.
Velasco, A. A., S. Hernandez, T. Parsons, and K. Pankow (2008), Global ubiquity of dynamic earthquake triggering, *Nat. Geosci.*, *1*, 375–379, doi:10.1038/ngeo204.
Wang, J. P., D. Huang, S.-C. Chang, and Y.-M. Wu (2014), New evidence and perspective to the Poisson process and earthquake temporal distribution from 55,000 events around Taiwan since 1900, *Nat. Hazard Rev.*, *15*(1), 38–47.
Woessner, J., and S. Wiemer (2005), Assessing the quality of earthquake catalogs: Estimating the magnitude of completeness and its uncertainty, *Bull. Seismol. Soc. Am.*, *2*, 684–698, doi:10.1785/012040007.
Zhuang, J., Y. Ogata, and D. Vere-Jones (2004), Analyzing earthquake clustering features by using stochastic reconstruction, *J. Geophys. Res.*, *109*, B05301, doi:10.1029/2003JB002879.